



UNIVERSITAT DE
BARCELONA

Understanding the link between chromatin structure, chromosome conformation and gene regulation

Diana Camila Buitrago Ospina



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – CompartirIgual 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – CompartirIgual 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-ShareAlike 4.0. Spain License.**

UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DOCTORAT EN BIOMEDICINA

Understanding the link between
chromatin structure, chromosome
conformation and gene regulation

Diana Camila Buitrago Ospina

September 2019

UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DOCTORAT EN BIOMEDICINA

Tesis realitzada en el grup de Modelatge Molecular i Bioinformàtica (MMB)

Institut de Recerca Biomèdica de Barcelona



INSTITUTE
FOR RESEARCH
IN BIOMEDICINE



UNIVERSITAT DE
BARCELONA

Understanding the link between chromatin structure, chromosome conformation and gene regulation

Memòria presentada per Diana Camila Buitrago Ospina per optar al grau de doctora per la Universitat de Barcelona

Tutor i Director

Modesto Orozco López

Doctorand

Diana Camila Buitrago Ospina

To Juan Camilo

To my parents Esperanza and Gustavo

Abstract

Understanding the connection between DNA organization in the nucleus, and cell functioning is one of the most intriguing problems in biology. Although many interdisciplinary efforts have been developed for this aim, the mechanisms of DNA folding in such a large scale are largely unknown. Therefore, the complexity of genome structure requires different techniques to tackle several resolution levels.

In this thesis, several scales of genome folding are studied using theoretical methods. First, we focus on the DNA sequence dependent properties which define the propensity of specific loci to be recognized by proteins, finding that the flexibility of specific DNA sequences might explain their prevalence in the genome.

DNA sequence dependent properties are also important to define the first layer of chromatin organization: the nucleosome. Physical descriptors of the DNA sequence combined with the propensity for transcription factor binding are highly informative on the location of nucleosome depleted regions, which guide the position of +1 and -last nucleosomes, the rest of nucleosomes in the gene body being placed by statistical phasing. There is a clear correlation between transcriptional activity and nucleosome phasing at gene body, the causal relationship is transcription→nucleosome organization rather than the opposite

A package for the comparative analysis of nucleosome organization was also developed in this thesis to quantitative predict changes in nucleosome organization occurring when perturbations are introduced to the cell.

Finally, we studied both the changes at the nucleosome level and at larger scale produced by the induction of DNA methylation on a natively unmethylated genome, developing a Hi-C based 3D model to gain insights into the chromatin rearrangements observed. We found very significant changes in chromatin structure induced by methylation, which are reflected in gene expression and cellular phenotype. Interestingly, these changes are found in a model organism that has not proteins prepared to recognize methylation, and accordingly can be assigned to intrinsic (not protein-mediated) effects of methylation.

Thesis organization

This thesis is a compilation of five works, three published and two in the process of publication, that study DNA and chromatin structure and its relation to gene regulation. They are presented following the level of resolution analyzed, rather than the chronological order of publication, starting from the structural properties of DNA sequences, following to the nucleosome organization and finally studying the 3D organization of the chromatin in the nucleus. *Chapter 1* starts with an introduction of the state-of-the-art about many of the aspects addressed in this thesis, as well as the general objectives proposed in this work. *Chapter 2* summarizes the methods used along the different projects, comprising the theoretical study of DNA physical properties as well as several next generation sequencing experiments, and the bioinformatics algorithms for their analysis. These methods allow the study of different genomic features such as nucleosome positioning, 3D genome organization, protein binding mapping and gene expression. *Chapters 3-6* present the results of this thesis as a compendium of five publications, each one preceded by a brief contextualization and summary of the main results. *Chapter 3* covers two publications concerning the importance of sequence dependent physical properties of the DNA on its flexibility and protein binding. *Chapter 4* presents a study of determinants of nucleosome positioning and its relationship with gene expression, combining information about intrinsic physical properties of the DNA with extrinsic features such as transcription factor binding. *Chapter 5* moves to the analysis of experimental data for nucleosome positioning, presenting a package developed not only to extensively analyze the nucleosome organization in a given experiment, but also to compare between different experimental conditions and to put the results in context with other genomic information. *Chapter 6* studies the effect of DNA methylation on chromatin structure at the nucleosome and whole-genome 3D levels, in engineered yeast to which DNA methyltransferases were transferred. Finally, *Chapter 7* contains a general discussion of the results presented in this work and the conclusions of this PhD thesis.

Acknowledgments

This work was possible thanks to the help and guidance of many people, starting from my supervisor, Prof. Modesto Orozco from whom I have learnt so much. All these years I have met many inspiring people in the lab. A special thanks to Fede who always believed in me and whose support, enthusiasm and encouragement were crucial to finish this work. I want to express my gratitude to Oscar and Pablo who were my mentors in the group; Juan Pablo, Jurgen, Ricard and Pau for all the contributions to this work; Alexandra who always found a moment to share some thoughts. I would like to thank Laia, Romina, Adam, Genis and Diego for their informatic support and hard work to produce a great virtual research environment, Marga for her assistance with all administrative tasks and all members of MMB group with who I crossed paths along these years.

I am also in debt with Isabelle who always had her door open to answer my questions and give some sense to my results, and the team at EBL, specially Rafael and Mireia who provided all the experimental data in this work.

I am also grateful to Simon Heath for receiving me in a short stay at his group at the beginning of my PhD, and for sharing his experience in statistics every time I needed; Marta for a wonderful collaboration and for bringing her motivation and interesting ideas to the lab.

Finally, this work would not have been possible without all my family. My deepest gratitude and love to my parents for their constant support and encouragement. Andre for always challenging me to think outside the box and to question every idea. Lina for being an inspiration and example of perseverance. Last but not least, thanks to my husband Juan Camilo for supporting me every day of my life and encouraging me to never give up, I have no words for expressing my gratitude to you.

Table of Contents

Abstract.....	v
Thesis Organization	vii
Acknowledgments	ix
List of Figures	xv
Abbreviations & Acronyms	xvii
PhD Advisor Report	xix
Chapter 1. Introduction	1
1.1 <i>Nucleosomes are the primary units of genome organization</i>	<i>1</i>
1.2 <i>Genome-wide nucleosome organization</i>	<i>3</i>
1.2.1 Sequence determinants of nucleosome positioning	4
1.2.2 Transcription regulation influences nucleosome positioning	5
1.2.3 Nucleosome architecture and DNA replication.....	8
1.2.4 ATP-dependent chromatin remodelers.....	9
1.2.5 Evidence of subnucleosomal structures	10
1.3 <i>Epigenetics dynamically modulates chromatin structure</i>	<i>10</i>
1.3.1 Histone post-translational modifications	10
1.3.2 DNA methylation.....	11
1.4 <i>Chromatin structure at higher level</i>	<i>13</i>
1.4.1 Nucleosome arrays form a second layer of chromatin organization ..	13
1.4.2 Higher level chromatin organization	14
1.4.3 TAD formation.....	16
1.4.4 Role of TADs in transcriptional regulation.....	16
1.5 <i>Chromatin organization in yeast.....</i>	<i>17</i>
1.5.1 Cell cycle chromatin dynamics.....	18
1.5.2 Chromatin organization under quiescence.....	20
1.5.3 Replication origins	20
1.5.4 tRNA genes	21
1.6 <i>Chromatin modeling.....</i>	<i>21</i>
1.6.1 Bottom-up chromatin models.....	22
1.6.2 Top-down chromatin models	23
<i>Bibliography for Chapter 1</i>	<i>26</i>
Objectives	35

Chapter 2. Methods.	37
2.1 <i>DNA physical properties</i>	37
2.2 <i>Transcription factor binding</i>	39
2.3 <i>Nucleosome positioning</i>	40
2.3.1 Studying nucleosome positioning <i>in vivo</i>	40
2.3.2 Mapping and noise filtering of the MNase signal	42
2.3.3 Nucleosome calling with nucleR	42
2.4 <i>Chromatin 3D structure</i>	43
2.4.1 Chromosome conformation capture	44
2.4.1.1 Hi-C	44
2.4.1.2 Capture Hi-C	45
2.4.1.3 Micro-C	45
2.4.2 Quantification of contact frequencies	45
2.4.2.1 Quality control	46
2.4.2.2 Mapping	46
2.4.2.3 Fragment-level filtering	47
2.4.2.4 Bin-level filtering and normalization	48
2.4.3 Identification of differential interactions	48
2.4.4 Visualization	49
2.5 <i>DNA methylation</i>	50
2.5.1 Whole-genome bisulfite sequencing	50
2.5.2 Nanopore sequencing	51
2.6 <i>ChIP-seq</i>	52
2.7 <i>RNA-seq</i>	53
<i>Bibliography for Chapter 2</i>	54
Chapter 3. Sequence dependent DNA flexibility and protein recognition	57
3.1 <i>Modulation of the helical properties of DNA: next-to-nearest neighbor effects and beyond (Publication 1)</i>	59
3.2 <i>Sequence selective protein-DNA recognition (Publication 2)</i>	74
<i>Bibliography for Chapter 3</i>	92
Chapter 4. Determinants of nucleosome architecture in yeast (Publication 3)	93
<i>Bibliography for Chapter 4</i>	110
Chapter 5. Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning (Publication 4)	117
<i>Bibliography for Chapter 5</i>	135
Chapter 6 . Impact of DNA methylation on 3D genome structure	137
<i>Bibliography Chapter 6</i>	174

Chapter 7 . General discussion and conclusions.....	175
7.1 <i>Sequence dependent DNA flexibility and protein recognition.....</i>	<i>175</i>
7.2 <i>Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning</i>	<i>177</i>
7.3 <i>Impact of DNA methylation on 3D genome structure.....</i>	<i>178</i>
Conclusions.....	180
Resumen.....	183
Introducción.....	185
Objetivos.....	188
Discusión general	190
Conclusiones.....	197
Annexes.....	201
I. <i>Modulation of the helical properties of DNA: next-to-nearest neighbor effects and beyond</i>	<i>201</i>
II. <i>Sequence selective protein-DNA recognition.....</i>	<i>222</i>
III. <i>Determinants of nucleosome architecture in yeast.....</i>	<i>250</i>
IV. <i>Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning</i>	<i>261</i>
V. <i>Impact of DNA methylation on 3D genome structure.....</i>	<i>279</i>

List of Figures

Figure 1.1. Nucleosome core particle from crystal structure	2
Figure 1.2. Nucleosome occupancy and positioning	3
Figure 1.3. Periodic positioning of A/T and C/G in the nucleosome dyad	5
Figure 1.4. Nucleosome free regions around TSS.....	6
Figure 1.5. Hierarchical chromatin organization.....	15
Figure 1.6. <i>Saccharomyces cerevisiae</i> chromatin structure.....	18
Figure 1.7. Chromatin organization throughout the cell cycle.....	19
Figure 1.8. Chromatin polymer modelling from epigenomic domains	22
Figure 1.9. Representation of the SBS model of chromatin	23
Figure 1.10. Example of results from a top-down chromatin model	24
Figure 2.1. Base pair step helical parameters representation.....	38
Figure 2.2. MNase-seq experimental procedure.....	41
Figure 2.3. Nucleosome positioning from MNase-seq data with nucleR.	43
Figure 2.4. Overview of the steps to perform a Hi-C experiment.	45
Figure 2.5. Mapping strategies implemented in TADbit.....	47
Figure 2.6. Fragment filtering in Hi-C data	48
Figure 2.7. Visualization of a Hi-C contact matrix in the 3D Genome Browser.....	50
Figure 2.8. Conversion of cytosines after treatment with bisulfite.	51
Figure 4.1. Nucleosome coverage around the -last nucleosome and the TTS..	94
Figure 4.2. Example of prediction of nucleosome coverage for a gene body.	95
Figure 4.3. Effect of transcription inhibition on nucleosome coverage.	96
Figure 5.1. Comparison of two nucleosome profiles using NucDyn.....	118
Figure 5.2. Example of results of nucleR and other nucleosome-related analyses	119
Figure 5.3. Visualization of Nucleosome Dynamics results in the MuGVRE.....	120

Abbreviations & Acronyms

3C	–	Chromosome conformation capture
3D	–	Three-dimensional
ChIP	–	Chromatin Immunoprecipitation
DNMT	–	DNA methyltransferase
ML	–	Machine Learning
MNase	–	Micrococcal nuclease
MD	–	Molecular Dynamics
NFR	–	Nucleosome Free Region
NRL	–	Nucleosome Repeat Length
PDB	–	Protein Data Bank
RE	–	Restriction enzyme
r.p.m	–	Reads per million
SMC	–	Structural maintenance of chromosomes
SPB	–	Spindle pole body
TET	–	Ten-eleven translocation
TFBS	–	Transcription Factor Binding Site
TSS	–	Transcription Start Site
TTS	–	Transcription Termination Site
WGBS	–	Whole Genome Bisulfite Sequencing

PhD Advisor Report

Publication 1

Alexandra Balaceanu, Diana Buitrago, Jurgen Walther, Adam Hospital, Pablo D. Dans. and Modesto Orozco. (2019). Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. Nucleic Acids Research, 47, 4418–4430. <https://doi.org/10.1093/nar/gkz255>.

The paper was published in Nucleic Acids Research, with an impact factor of 11.147 (Q1). D. Buitrago performed the genome wide characterization of the tetramer in the genome of several organisms as well as in cancer data sets containing mutational signatures.

Publication 2

Federica Battistini, Adam Hospital, Diana Buitrago, Diego Gallego, Pablo D. Dans, Josep Lluís Gelpí, and Modesto Orozco. (2019). How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. Journal of Molecular Biology S0022-2836(19)30451-6. <https://doi.org/10.1016/j.jmb.2019.07.021>.

The paper was published in Journal of Molecular Biology, with an impact factor of 5.067 (Q1). D. Buitrago performed all the statistical analyses in the manuscript to test the difference between experimental and MD simulation helical parameters.

Publication 3

Diana Buitrago^{*}, Mireia Labrador^{*}, Pau De Jorge, Federica Battistini, Isabelle Brun Heath and Modesto Orozco. The interplay between periodicity, DNA physical properties and effector binding define nucleosome architecture in yeast (*in preparation*).

D. Buitrago is the first co-author of the paper. She performed all the bioinformatic and machine learning analyses in the paper. The paper is in its final stage of preparation and will be submitted after obtaining the sequencing results and analyzing the ChIP-seq experiment for RNA Polymerase, to confirm one of the hypotheses stated in the paper.

^{*} Equally contributing authors

Publication 4

Diana Buitrago^{*}, Laia Codó^{*}, Ricard Illa, Pau de Jorge, Federica Battistini, Oscar Flores, Genis Bayarri, Romina Royo, Marc Del Pino, Simon Heath, Adam Hospital, Josep Lluís Gelpí, Isabelle Brun Heath and Modesto Orozco. (2019). Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning. Nucleic Acids Research, gkz759. <https://doi.org/10.1093/nar/gkz759>.

The paper was published in Nucleic Acids Research, with an impact factor of 11.147 (Q1). D. Buitrago is first co-author and developed the code, analyzed the examples and performed the benchmark comparing Nucleosome Dynamics with other previously published algorithms.

^{*} Equally contributing authors

Publication 5

Diana Buitrago^{*}, Mireia Labrador^{*}, Simon Heath, Juan Pablo Arcon, Rafael Lema, Oscar Flores, Anna Esteve-Codina, Julie Blanc, David Bellido, Marta Gut, Ivo Gut, Pablo D. Dans, Isabelle Brun Heath, Modesto Orozco. Impact of DNA methylation on 3D genome structure. (*in preparation*)

D. Buitrago is first co-author of the paper. She performed most of the bioinformatic analyses in the paper and participated in the development of the 3D chromatin model. The paper is in its final stage of preparation and will be submitted by the time of thesis defense.

^{*} Equally contributing authors

Prof. Modesto Orozco

Modesto.orozco@irbbarcelona.org

Chapter 1. Introduction

DNA is a long molecule that under physiological conditions forms a complementary right-handed duplex containing the genetic information necessary to build life. Although the human DNA fiber is about two meters long, it is packed tightly to fit inside the small space defined by the cell nucleus with a diameter of approximately 10 micrometers [1]. The DNA compaction is aided by proteins that guide DNA folding inside the nucleus of eukaryotic cells. The complex of DNA and proteins inside the nucleus is known as chromatin. Many experimental evidences [2]–[4] demonstrate that DNA packing inside the nuclei is not random, as the accessibility to DNA of genome regulators must be preserved, ensuring the correct function of processes such as transcription, replication, and DNA repair. Other evidences have shown that this organization is dynamic and undergoes different rearrangements along several cellular processes such as differentiation [2], cell cycle progression [5] or damage response [6].

1.1 Nucleosomes are the primary units of genome organization

The fundamental unit of DNA compaction in eukaryotes is the nucleosome. A canonical nucleosome is formed by ~147 base pairs (bp) of double-stranded DNA that coil in approximately 1.65 super helical turns around a core of histone proteins, which contain two copies of each histone H2A, H2B, H3 and H4. X-ray crystal structures of the nucleosome [7], [8] revealed (**Figure 1.1**) that histone proteins are

formed by a globular domain constituting the nucleosome core with a disk-like shape (approximately 10 nm of diameter and 5.5 nm of height) and N-terminal histone tails that are relatively unstructured and highly flexible. Moreover, they can undergo post-translational modifications (methylation, acetylation, ubiquitination) altering chromatin accessibility [9].

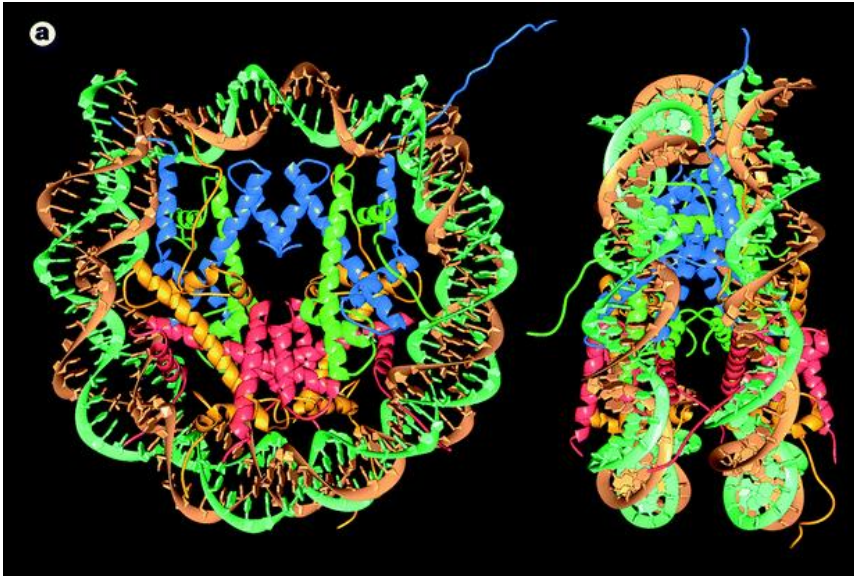


Figure 1.1. Nucleosome core particle derived from crystal structure at 2.8 Å resolution. Histone ribbon traces colored in blue (H3), green (H4), yellow (H2A) and red (H2B). Adapted from [7].

The central bases of the nucleosomal DNA coincide with a pseudo 2-fold symmetry axis, the dyad axis [10]. The high curvature of the DNA in the nucleosome requires significant bending energy [11]. The structure is stabilized by positively charged histones in complex with negatively charged DNA backbone that form interactions every 10 bp (formed by salt-bridges, hydrogen bonds and hydrophobic contacts [10]) and interactions between the histones, forming H3-H4 and H2A-H2B dimers. The DNA wraps around the histones are parallel except in the entry/exit of the DNA to the nucleosome, where an additional histone binds, known as linker histone (H1 or H5), present in higher eukaryotes. Linker histones have an important role in interactions between nucleosomes and, hence, in folding of the nucleosomes in space and chromatin compaction [12].

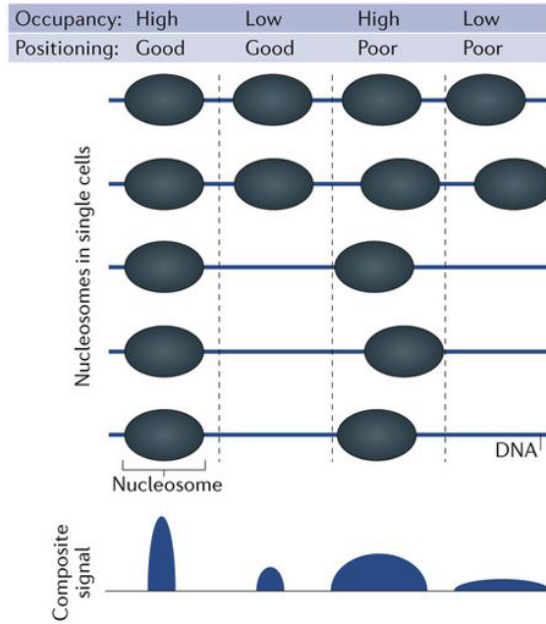


Figure 1.2. Nucleosome occupancy and positioning. A nucleosome in a pool of cells can be characterized by the relative number of cells that contain it (occupancy) and the variability between cells in the sequence position (positioning). Adapted from [13].

1.2 Genome-wide nucleosome organization

Nucleosome positions along genomes *in vivo* have been determined using several experimental protocols, such as FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) [14], ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) [15] and MNase-seq (Micrococcal Nuclease digestion followed by high throughput sequencing) [16]. The latter is the most widely used technique and provides detailed information on nucleosomal architecture. It is based on cross-linking nucleosomal DNA and histones using formaldehyde prior to treatment with MNase, which cleaves the linker fragments. These experiments contain information from a population of cells, therefore nucleosome profiles can be noisy [17] and are typically characterized by two important properties: occupancy and positioning (**Figure 1.2**). The first is related to the percentage of cells in an experiment that contains a given nucleosome, the latter denotes the variability in its genomic position among all the cells. A nucleosome is called well-positioned (W) when it is present in a large percentage of the cells, and the fragments from different cells present low

variability with respect to the genomic position. When a nucleosome has low coverage and/or large positioning variability, it is called fuzzy (F) [17].

Nucleosome organization along the linear DNA sequence is not random and it has been related to different cellular processes such as transcription and replication [13]. Moreover, it is highly dynamic in space and time, and influenced by several factors such as: (i) the local context determined by sequence-dependent properties (*cis*-factors), (ii) protein complexes that interact with the DNA and can compete with nucleosomes (*trans*-factors), such as transcription factors [18], replication machinery [13] or ATP-dependent remodelers that can slide or evict nucleosomes (partially or totally) [19], and (iii) the effect of neighboring nucleosomes that impose steric constraints for nucleosome positioning [20]. A summary of these factors is presented in the remaining of this section.

1.2.1 Sequence determinants of nucleosome positioning

As explained above, the B-DNA conformation is highly distorted as it is wrapped around the histone core. Since DNA sequences are characterized by different physical properties depending on the bp composition, it is expected that some sequences are more favorable to form nucleosomes [21], [22], [23]. Efforts to determine the sequence contribution to position nucleosomes have been performed *in vivo* and *in vitro* [20], [19], [24].

High resolution nucleosome maps in budding yeast have revealed that nucleosome formation is favored in GC-rich sequences whereas poly(dA:dT) sequences tend to be nucleosome depleted [20]. Alignment of thousands of well positioned nucleosomes showed a preferential periodic pattern (**Figure 1.3**) of AA, TT, TA and AT steps every DNA turn (10 bp) offset by 5 bp of another periodic pattern of G/C dinucleotides [13]. This is related with the thermodynamically favoring of AA, TT and TA to expand the DNA major groove and CG to contract it [22], [23].

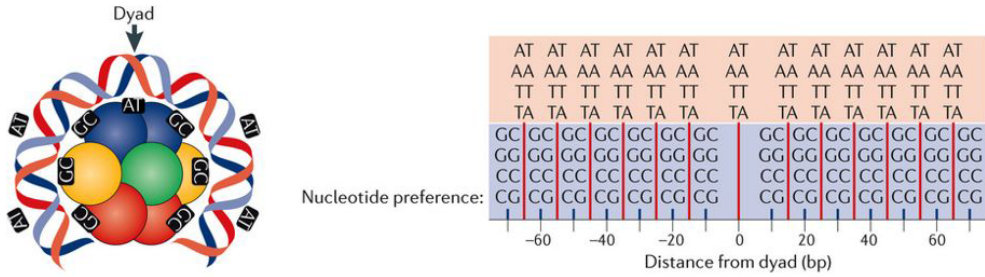


Figure 1.3. Preferential periodic positioning of A/T and C/G di-nucleotides relative to the nucleosome dyad. Nucleosome dyad tends to be enriched in A/T di-nucleotides. A 10 bp periodic pattern of A/T di-nucleotides is preferentially found, offset 5 bp by C/G di-nucleotides. Adapted from [13].

The observation of this preferential sequence positioning pattern for nucleosomes led to a large number of models that predict nucleosome positioning from sequence composition [24]. Some of these models (reviewed in [20]) compute the elastic energies associated with DNA bending characteristic to each genomic sequence, others use the periodicity of favorable or unfavorable dinucleotides or the frequency of different k-mers in nucleosomal sequences, coupled with machine learning (ML) algorithms to predict nucleosome positions genome-wide. *In vitro* predictions are more accurate, since the effect of interacting proteins or chromatin remodelers is removed [22]. However, the predictive ability of classical sequence-based models is moderate *in vivo* as the *trans*-acting factors are also important for nucleosome positioning [25].

1.2.2 Transcription regulation influences nucleosome positioning

Genome-wide nucleosome positioning studies have revealed the presence of a nucleosome free region (NFR) around promoters [13], [26], surrounded by two strongly positioned nucleosomes referred to as -1 and +1 (the nucleosomes immediately upstream or downstream the transcription starting site, TSS, respectively). Genes can be characterized by these NFRs surrounding nucleosomes, taking into account their positioning (F or W) and the distance between their dyads (forming open or close conformations) [17].

Promoters of active genes tend to be associated with open NFRs (**Figure 1.4**), where transcription factors or RNA polymerases can be bound [27]–[29]. On the contrary,

genes with low transcription levels tend to present narrower linkers where transcription factor (TF) binding site might be occluded by the nucleosomes [30]. However, this tendency is not present in all genes, and in fact some TFs can bind nucleosomal DNA either at the nucleosome exit or dyad, at motifs formed by the two parallel DNA chains that surround the histones, or at 10 bp periodic motifs favored by the conformation of the DNA in the nucleosome [18], [31]. Moreover, nucleosome-bound TFs can have opposite roles in gene activation (TF binding leads to nucleosome dissociation) or repression (nucleosome stability is increased upon TF binding, for instance in T-box TFs) [18], [32].

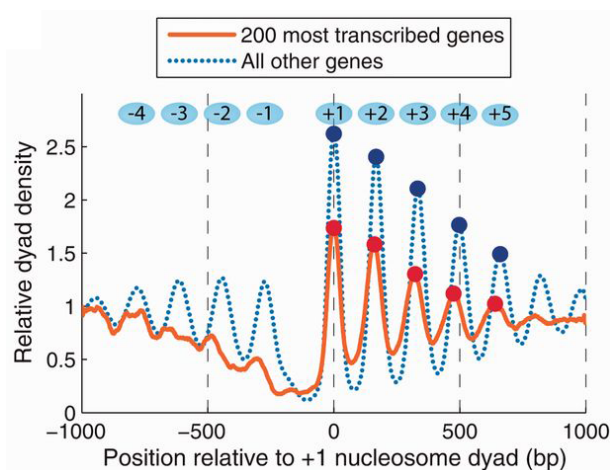


Figure 1.4. Nucleosome free regions around TSS of genes according to transcription levels. Highly transcribed genes present wider NFRs. Adapted from [30].

The nucleosome architecture around TSS could be both cause or consequence of the transcriptional activity in a given cell condition. During transcription, RNA polymerase II (RNA pol II) passage seems to require partial disruption of the DNA-histone contacts within promoters and coding regions but rapid reestablishment of the chromatin organization occurs after polymerase passage [13]. In particular, loss of -1 nucleosome due to RNA pol II binding has been reported as well as shortening of the NFR due to increased occupancy at nucleosome -1 under polymerase activity loss [33]. At +1 nucleosome, histone variant H2A.Z might facilitate the passage of RNA polymerase [34]. Along gene bodies, transcriptionally inhibited cells exhibit low nucleosome positioning [35]. Recent cryoEM studies [36], [37] have determined

the structure of RNA pol II passage along nucleosomal DNA, showing how transcription elongation factors can accommodate to the nucleosome core, and how the DNA is peeled from the nucleosome core. Although these studies are preliminary, they set the path for further investigation on transcriptional elongation along chromatin.

Further evidence about the relationship between transcription and nucleosome positioning was presented in studies showing how nucleosome shifts or evictions might appear related to gene activation. Moreover, binding of pre-initiation complex in order to activate transcription might require a specific nucleosome position around the TSS, possibly to allow accessibility to the TF binding site [35]. For instance, under heat-shock, *Saccharomyces cerevisiae* ribosomal protein promoters are downregulated, accompanied with an eviction of DNA binding factors and an upstream shift of the +1 nucleosome [38]. Shifts and evictions of nucleosomes around gene promoters, related to transcriptional activation, have also been reported in human cells [16], [29].

The strong positioning of the +1 nucleosome has led to development of models that hypothesize the presence of barriers at promoters imposing a periodic organization of the downstream nucleosomes [13]. Several models [24] have proposed the prediction of nucleosome profiles using an emitting signal from these barriers and positioning nucleosomes at an average distance between nucleosomes, known as nucleosome repeat length (NRL), which varies between cell types and chromatin states [16], [29], [39], [40]. Moreover, the periodicity in nucleosome spacing varies along the chromosomes [41], but there is no consensus regarding the correlation between nucleosome periodicity and transcription level of the corresponding genes. From MNase-seq studies, including nucleosomal fragments from a population of cells, contradictory results have been found, since periodically positioned nucleosomes have been found both to promote and inhibit transcription initiation [13]. On the contrary, studies analyzing single-cell nucleosome profiles found that silent genes display periodic nucleosome arrays (although with large variation of the exact genomic positioning between cells) while nucleosomes in actively transcribed genes are less periodic, but their position is more conserved among the different cells.

These results were found using single cell MNase-seq [42] and Array-seq [43], a technique sequencing di-, tri-, tetra-nucleosomes to extract nucleosome spacing in individual DNA molecules. Hence, the study of nucleosome periodicity should consider possible artefacts caused by cancelling effects between the different cells in population-based MNase-seq data.

The effect of the transcription termination site, TTS, as an emitting signal for nucleosome positioning is less clear. There is evidence supporting the existence of a barrier for nucleosome positioning from the 3' gene end in *S. cerevisiae* [44] but it can be influenced by the proximity between its TTS and the TSS from downstream genes [45]. However, in human cells, where intergenic regions are wider, presence of poly(A) sequences at 3' gene ends suggests that the nucleosome positioning barriers at TTS are not an artefact of neighboring genes [46].

1.2.3 Nucleosome architecture and DNA replication

Replication is initiated at specific locations of eukaryotic genomes, where the origins recognition complex (ORC) binds to some consensus sequences and recruits factors required for DNA replication [13]. Replication origins have different firing times: those activated shortly after entrance to S-phase called "early" origins and those active at the end of S-phase called "late" origins [47]. In *S. cerevisiae*, replication origins, also known as autonomously replicating sequences (ARS) have been broadly identified [48]. Budding yeast ARS consensus sequences (ACS) display two different nucleosome architectures depending on their activity. Active ACS coincide with long NFRs surrounded by strongly positioned nucleosomes, that are enriched in histone variant H2A.Z [13]. On the other hand, functionally inactive ACS are partially covered by nucleosomes [49]. Replication origins in mammalian cells have also been found to coincide with NFRs although a combination of several factors determines the final origin profile in different cell types [50].

As replication progresses, nucleosome structure is perturbed, since the replication complexes need access to single stranded template DNA [47]. Rapid restoration of chromatin organization must take place now in two DNA molecules, which implies

two octamers of histones are required per each original nucleosome. Moreover, histone variants and modifications must be faithfully preserved. Three models of histone H3-H4 inheritance have been proposed [13]: (i) conservative, where one of the daughter DNA duplexes remains bound to the original histones and the other is assembled with a new set of histones, (ii) semi-conservative, where each daughter keeps half of the original histone content, and (iii) dispersive, which is a mix of the previous two, depending on the histone variant composition.

1.2.4 ATP-dependent chromatin remodelers

We have seen that transcription and replication might require nucleosome displacements or disassembly. In order to achieve this, ATP-dependent chromatin remodeling complexes use the energy from ATP hydrolysis to reposition nucleosomes [19]. Transcription, DNA replication and DNA repair can also require histone turnover, mediated by chromatin remodelers. Although turnover rate is higher in active promoters, enhancers and origins of replication, opposing evidence shows that nucleosomes at highly transcribed regions are maintained, suggesting that nucleosome turnover upon gene activation could be only partial [13]. Moreover, there is evidence that binding of remodelers to promoters contributes to define a strong NFR and phased arrays of nucleosomes [35], [51].

Different types of ATP-dependent chromatin remodelers have been described, depending on the catalytic subunit of the remodeling enzymes. A summary of their main roles [19], [52] is presented in the following:

- SWI/SNF is highly conserved in eukaryotic cells, implicated in regulation of stress response.
- RSC is highly abundant and required for cell viability. It can produce nucleosome shifts at promoters producing wide NFRs in active genes.
- CHD is involved in DNA replication and repair. It participates in the regular spacing of nucleosomes.

- ISWI is a highly conserved complex important in transcription and DNA replication. Repositions nucleosomes similarly to SWI/SNF and can also affect nucleosome spacing.
- INO80 is involved in transcription, DNA repair and replication. Many eukaryotes share a conserved core of subunits from this complex, but other subunits have largely diverged through evolution.

1.2.5 Evidence of subnucleosomal structures

Besides from the canonical histone octamer, the existence of subnucleosomal structures has been reported. For instance, centromeric nucleosomes in *Drosophila melanogaster* were reported to contain only one copy of each histone protein centromeric H3 (cenH3), H4, H2A, H2B forming an “hemisome” structure [53], while *S. cerevisiae* centromeric nucleosomes were found to be hexamers formed by Cse4 (instead of H3), H4 and Smc3 (in place of H2A and H2B), named hexasome [54].

Hexasomes at +1 nucleosomes in *S. cerevisiae* with an unbalanced composition of H3-H4 histones were reported, which coincide with regions of high histone turnover [55]. Moreover, as explained before, RNA polymerase passage can induce partial unwrapping of the nucleosomal DNA leading to opening or dissociation of H2A-H2B dimers, and therefore to nucleosomes with lower number of histones. Additionally, evidence for the existence of half-nucleosomes linking DNA replication with H3-H4 tetramers has been reported [13].

1.3 Epigenetics dynamically modulates chromatin structure

1.3.1 Histone post-translational modifications

Histone tails largely contribute to dynamics of the chromatin structure related to gene transcriptional regulation [9]. Tails interact with the DNA influencing nucleosome stability and recruitment of regulatory proteins [56] and they also interact between nucleosomes, modulating higher order structures[57]. Histone tails

can be subject to several post-translational modifications such as methylation, acetylation, phosphorylation and ubiquitination. More condensed groups of nucleosomes are observed in heterochromatic regions generally associated with H3K27me3 and H3K9me3/2 and less compact and more accessible in euchromatin, marked by H3K4 methylation lysine acetylation and preferentially located in the interior part of the nucleus [58].

Histone acetylation has generally been related to higher gene expression by decreasing chromatin compaction [57]. Histone acetyltransferases add acetyl groups to lysine residues, neutralizing the positive charge of histone tails and therefore reducing its affinity to the negatively charged DNA. An opposite role is attributed to histone deacetylases which, by removing acetyl groups from histone lysine residues, make chromatin more compact and therefore have a repressive role [59].

Histone methylation has a dual role in transcription activation or repression, depending on the target residue and the number of methyl groups added [56]. Although methylation of lysines and arginines does not alter the electrostatics of DNA-histone interactions, as it occurs with acetylation, its effect on activation or repression is related to different regulatory proteins that are recruited depending on the precise modification [9].

The effect of histone phosphorylation and ubiquitination is coupled to other histone modifications, defining an interplay between them where the presence of one modified residue can induce the epigenetic modification of another [9]. Histone phosphorylation is related to several processes such as DNA repair, chromatin compaction in mitosis and regulation of gene expression [56]. Ubiquitination is also related to different activities, such as DNA damage signaling and transcriptional activation (both activation or repression, depending on the target residue) [60].

1.3.2 DNA methylation

Another dynamic epigenetic modification that is correlated with gene silencing and chromatin conformation in bacteria, plants and mammalian cells is DNA methylation. This modification it is not present in all eukaryotes; it is absent in several

model organisms such as *S. pombe*, *S. cerevisiae* and *C. elegans*, and barely detectable in *D. melanogaster* [61]. DNA methylation occurs preferentially at cytosines, mainly at CpG steps, that are covalently modified. CpG steps are under-represented in the genome of complex organisms but enriched in approximately 60% of human promoters (mainly at CpG islands), suggesting a role for CpG methylation in gene regulation that might be coupled to the unusual conformational properties of CpG steps [62].

Methylation is established by DNA methyltransferases (DNMTs) whereas TET proteins are responsible for removal of the methyl groups. DNMT3a and DNMT3b are responsible for de-novo CpG methylation in both strands while DNMT1 participates in maintenance of CpG methylation after DNA replication [63]. Additionally, several methyl CpG binding domain (MDB) proteins are readers of methylated DNA and can modulate gene expression through changes in DNA accessibility and recruitment of different protein complexes [64]–[66] in higher eukaryotes.

In early stages of mammalian development, DNA methylation patterns are established, with most CpGs methylated except those located at CpG islands [63]. Upon differentiation, CpG islands in promoters of housekeeping genes remain unmethylated, but genes that are inactivated at a particular developmental stage get de-novo methylation [67] whereas other promoters and regulatory regions are demethylated [63]. Research has suggested that DNA methylation does not lead to gene repression, however it maintains the gene at inactive state [68] whereas demethylation can re-activate its expression [69].

Alterations in DNMTs have been linked to important effects in gene regulation associated to diseases and cell viability. For instance, mutations in DNMT3b are implicated in Immunodeficiency, Centromere instability and Facial anomalies (ICF) syndrome [70], while those in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy [71] and DNMT3a mutations are present in acute myeloid leukemia patients [72]. Furthermore, changes in DNA methylation patterns have been associated with many different types of cancer in humans [73], [74], related to

activation of oncogenes or repression of genes involved in DNA damage response [75], [76].

It is still unclear whether the effect of DNA methylation on gene expression is direct, or it is correlated with chromatin structure [77]. Theoretical analysis of the physical properties of DNA revealed that while CpG steps are very flexible [62], mCpG are stiffer and harder to bend, and have lower ability to circularize and to form nucleosomes [23], [78], [79]. However, *in vivo* studies on mammals and plants are contradictory, with some suggesting that methylation occurs preferentially on nucleosomal DNA [80], [81] and others concluding the opposite [82], [83]. NOME-seq experiments have shown that DNA methylation and nucleosome occupancy were strongly anti-correlated surrounding CCCTC-binding factor (CTCF) sites, but at promoters the correlation seemed to be less clear [84].

1.4 Chromatin structure at higher level

1.4.1 Nucleosome arrays form a second layer of chromatin organization

Nucleosomes are separated by fragments of DNA called “linkers”, between 10 and 100 bp long depending on the cell type and transcriptional state of the region [85]. Nucleosomes are connected by linkers in a beads-on-a-string array as detected initially by electron microscopy [86], [87]. Different experimental approaches have tried to elucidate their 3D folding. Early electron microscopy studies suggested that nucleosomes fold into a regular 30-nanometer fibers, but different folding motifs were identified, probably as a consequence of differences in experimental conditions [88]: one-start solenoid model [89], two-start helical ribbon model [90], and two-start crossed-linker model with left-handed double-helical symmetry [91]. Posteriorly, analysis of *in vitro* reconstituted nucleosomes also derived different folding patterns such as the zig-zag model [92]. However, *in vivo* studies questioned the existence of the regular 30 nm fiber and instead found evidence for random and irregular nucleosome arrangement [93], [94]. In recent years, with the advances in super-resolution microscopy, it has been observed that the chromatin fiber is not a regular

structure, but rather it is formed by groups with varying sizes that can be cell type specific [95], [96].

As pointed out in a recent review [88], the lack of consensus on a common folding motif for nucleosomes among different experiments can be due to differences in experimental conditions (for instance chromatin folding is dependent on salt concentration), but also to intrinsic variability of DNA sequences, epigenetic modifications, histone variants, or the effect of linker histones (H1). Chromatin conformation capture (3C) techniques (see Methods for detailed explanation) have provided very valuable information on the arrangement of the nucleosome fiber in the nucleus. Particularly, Hi-C experiments count the frequency of interaction between pairs of loci genome-wide, and a variant of the 3C technology named Micro-C [97], [98] (see Methods), allows the study of contact frequencies at the nucleosome resolution. These studies have been performed in *S. cerevisiae* and *S. pombe*, revealing the presence of self-associating domains that span 1 to 5 genes, usually separated by promoters of highly transcribed genes. They also found evidence for structural tri or tetra-nucleosome motifs. Another study using ionizing radiation-induced spatially correlated cleavage of DNA with sequencing (RICC-seq) found evidence for zig-zag nucleosome arrays in heterochromatic regions and solenoid structures in open chromatin [99]. Recently, Hi-C with nucleosome orientation (Hi-CO) [58], a method combining Hi-C with simulated annealing molecular dynamics, proposed alpha-tetrahedron and beta-rhombus tetra-nucleosome motifs, occurring preferentially at gene bodies and promoter regions, respectively.

1.4.2 Higher level chromatin organization

Hi-C [100], (see Methods) allows the interrogation of chromatin structure genome-wide at kilo base scale, revealing that chromosomes fold hierarchically in the nuclear space during interphase [101], [102], as illustrated in **Figure 1.5**. At the whole nucleus level, Hi-C experiments showed the segregation into chromosome territories, since contact frequencies between regions in the same chromosome (*cis*-contacts) are larger than between regions in different chromosomes (*trans*-contacts), as had been previously observed by FISH experiments in interphase nuclei [103]. At the mega

base scale, contact frequencies between loci have revealed the separation of A/B compartments, corresponding to actively transcribed euchromatin and repressed heterochromatin, respectively [100], the latter being preferentially attached to the nuclear lamina [104].

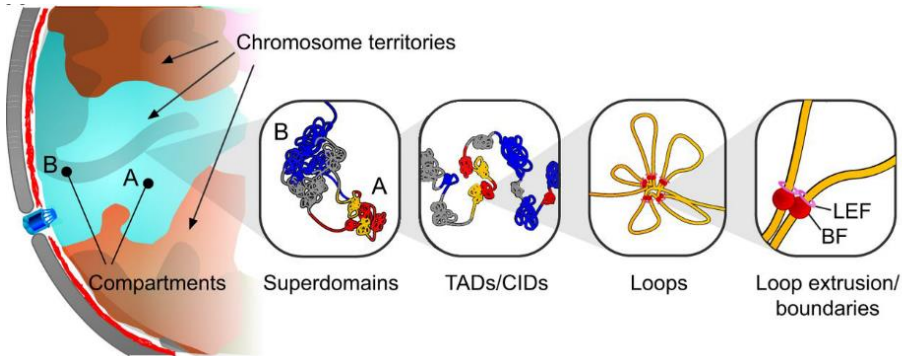


Figure 1.5. Hierarchical chromatin organization. Adapted from [105]

At finer scale, chromosomes are organized into topologically associated domains (TADs), regions of the genome with high self-interaction, insulated from regions of neighboring domains [106]. TADs might reflect the presence of gene loops that enforce promoter directionality [107] or loops formed to bring in proximity regulatory elements which can be separated by a large genomic distance, such as enhancers and their target sites, or co-regulated genes, [105], [108]. Loops have also been observed in mammalian genomes using ligation-free methods such as Genome Architecture Mapping (GAM) [109], that also reveals the abundance of three-way contacts between highly transcribed regions or super-enhancer loci. In smaller genomes such as *S. cerevisiae*, although Hi-C experiments initially failed to detect the presence of TADs [110], Micro-C allowed the detection of chromosomal interaction domains (CIDs), containing similar number of genes as a TAD [97], [98].

The link between chromatin structure and epigenetic states has also been studied. TADs tend to be formed by regions displaying similar accessibility and usually coincide with segmentation of chromatin by epigenetic profiles [108], [111]. Nonetheless, the effect of epigenetic marks, such as DNA methylation, on chromatin structure is still unclear. A/B compartments in many cell types can be

computationally predicted by DNA methylation profiles [112]. However, a recent study showed that, although the establishment of A/B compartments defines DNA methylation patterns in cardiac myocytes, alteration of DNA methylation signatures does not have an impact on chromatin compartmentalization or TAD formation [113].

1.4.3 TAD formation

TAD borders in mammalian cells strongly colocalize with CTCF target sites [114]. A proposed mechanism for TAD formation in interphase involves the role of loop extrusion factors (LEFs), for instance cohesin, that are loaded into DNA and extrude it through their ring-shaped structure [115], [116]. The loop formation continues until the LEF finds either another LEF or a boundary factor (BF), for instance an insulator CTCF in convergent orientation at the loop boundaries. Loops are not stable unities but rather as LEFs dissociate from chromatin the contacts between TAD boundaries can be lost and hence not detected in all cells of a population Hi-C experiment.

The loop extrusion model is supported by computer models that have shown that in metazoans, interphase domain formation requires cohesin-dependent looping [115], [116]. Moreover, reorganization in the TAD structure is observed upon depletion of cohesin or its loading factors [3], [117]. Either deleting CTCF binding motif or reversing its orientation can increase contacts between the two surrounding TADs, hence globally loosing insulation between TADs [106], [118].

1.4.4 Role of TADs in transcriptional regulation

As described above, TAD boundaries are enriched in bound CTCF in mammals, but in organisms such as *D. melanogaster*, the role of CTCF in TAD insulation is lower, and some organisms, for instance *S. cerevisiae*, do not have this protein (or a homolog). In those organisms, TAD or CID boundaries have been associated to promoters of actively transcribed genes and are typically bound by RSC remodeling complex [97], [119]. Moreover, mammalian TADs not only are enriched in CTCF

sites, but also coincide with promoters of housekeeping genes and open-chromatin marks [111].

Specific defects in genome folding have been related to failures in genome regulation leading to diseases and cancer [118], [120]. TADs seem to favor enhancer promoter interactions or promote co-localization of functionally related genes and hence are related to transcription activation [4], [121]. Additionally, chromatin compaction at gene level, obtained from Micro-C contact frequencies [97], is anticorrelated with transcriptional activity. Altogether, these results show wide evidence suggesting that TADs have an active role in transcription regulation. The converse relation, the effect of transcription on TAD formation, has also been studied. Although transcription alteration in *D. melanogaster* has shown effects in domain segregation, activation of a single gene does not create a TAD boundary in mammalian cells, suggesting that in the latter case the effect of CTCF might be stronger[122].

1.5 Chromatin organization in yeast

The first Hi-C studies in budding yeast revealed several features of its chromatin organization in asynchronous populations of cells, some of them confirming observations from microscopy experiments. Like other eukaryotes, contact frequencies are lower between chromosomes than within chromosomes and exponentially decrease with genomic distance [110]. On the other hand, longer chromosomes tend to have less interactions with other chromosomes. However, *S. cerevisiae* presents differences in chromatin folding with respect to higher eukaryotes that are enumerated below (**Figure 1.6**).

First, centromeres are clustered at the spindle pole body (SPB) through kinetochore microtubule attachment and have low interaction frequency with regions further in the same chromosome, presenting a Rabl-like organization, as observed by imaging and 3C studies [98], [110], [124].

Second, telomeres tend to localize towards the nuclear membrane and interact more frequently than expected considering their genomic separation. Several clusters of

telomers are observed within each single cell, preferentially formed by arms of similar length [125], [126].

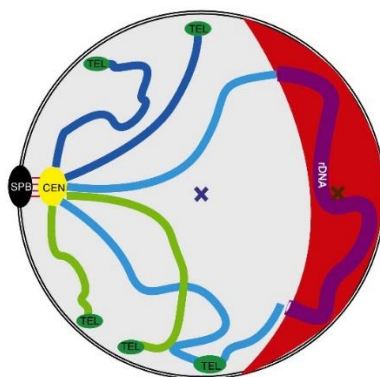


Figure 1.6. *Saccharomyces cerevisiae* chromatin structure in interphase. Centromeres are clustered attached to the SPB, telomeres preferentially located at nuclear periphery and rDNA segregated from the rest of chromatin at the nucleolus, opposite the centromere cluster. Adapted from Wang et al. 2015 [123].

Third, ribosomal DNA (rDNA) confined at the nucleolus is located in opposite side of the centromere cluster in interphase [110]. This region splits chrXII into two lowly interacting domains. Nucleolus volume in exponentially growing cells is approximately the third of the total nucleus volume [123]

These features are globally preserved in the two matting types in yeast: **a** and α , which are determined by the MAT locus on chromosome III. The only difference in chromosome folding between the two matting types occurs precisely in chrIII, where MAT locus is in contact with the heterochromatic locus HML in MAT**a** cells but not in MAT α cells. Moreover, a single loci, the recombination enhancer, is determinant in the reorganization of the whole chromosome [127].

1.5.1 Cell cycle chromatin dynamics

Along the cell cycle, the global patterns of budding yeast chromatin are preserved to some extent, although the intensity of centromere clustering, intra/inter contact ratio, rDNA compaction and nucleus sphericity suffer some variations. Hi-C experiments in cells arrested at different points along the cell cycle revealed

progressive increase in chromosome compaction between G1 and M [5], [128]. Structural maintenance of chromosomes (SMC) complexes are essential in mitotic chromosome condensation as well as chromatin structure in interphase [129]. By disruption of cohesin activity it was shown, both experimentally and through computer models, that the increased compaction achieved in M is dependent on cohesin but not condensin [128].

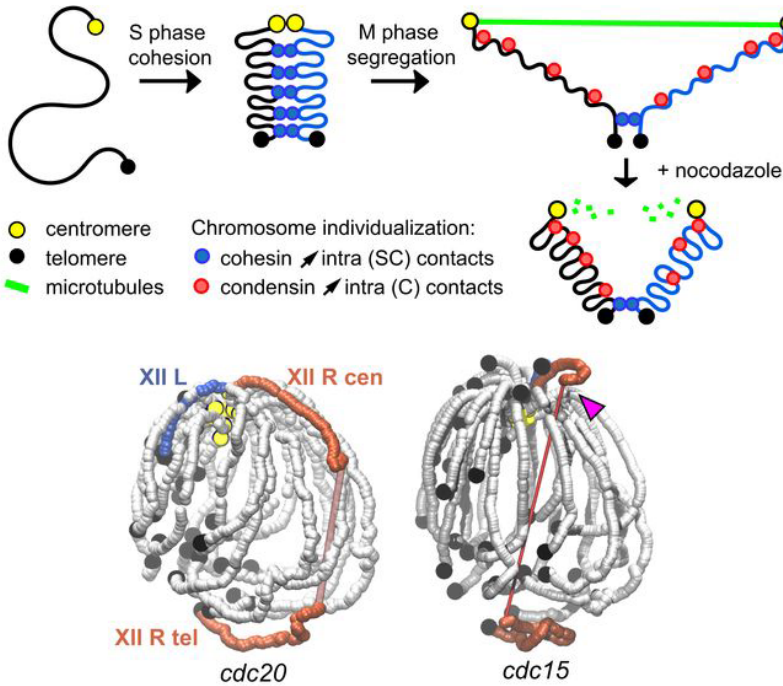


Figure 1.7. Chromatin organization throughout the cell cycle in *Saccharomyces cerevisiae*. Increase of intra-contacts at S phase mediated by cohesin leads to chromosome individualization. During M phase chromosomes are more elongated due to spindle elongation and condensin loading. Bottom: chromatin structure in *cdc20*- and *cdc15*-arrested *S. cerevisiae* cells, which correspond to metaphase and anaphase, respectively. Chromosome XII is highlighted in orange (right arm) and blue (left arm). The right arm folds into a loop in anaphase (pink arrowhead). Adapted from [5].

Progressing further into the cell cycle, in the transition from metaphase to anaphase, a third state of chromatin organization is observed. The structure resembles a polymer brush (**Figure 1.7**), with increased short-range contact frequency and decrease in long-range contacts. In anaphase, a strong loop between the upstream flanking region of rDNA and centromere is formed, and rDNA presents a stretched

conformation. Condensin participates in the formation of this loop, as well as in the increase in centromere contacts (also observed in other stages of the cell cycle, [128]).

1.5.2 Chromatin organization under quiescence

Under nutrient starvation, cells exit mitosis and enter quiescent state, referred to the G zero phase (G0), which produces changes in chromatin organization. Glucose starvation produces chromatin condensation, as observed by the reduction of nuclear volume in fluorescence microscopy [130] and the changes in chromatin contacts in Hi-C (increase in intra-chromosomal long-range contacts as well as contacts between the centromeric region and the rest of the chromosome, and decrease in inter-centromeric contacts, [131]).

rDNA acts as a barrier separating chrXII into two domains with very low interaction frequencies, but since it is more compact in quiescence due to ribosome biogenesis [132], the changes in long range contacts are stronger.

Additionally, telomeres, which form groups attached to the nuclear envelope in exponentially growing cells, form a single cluster in quiescent cells that is located in the center of the nucleus [131], [133].

1.5.3 Replication origins

Super-resolution microscopy revealed that replication origins are grouped into discrete points inside the cell nucleus, with high variability between different cells, forming foci of ARS that start replication simultaneously [134]. Mapping of ARS in *S. cerevisiae* revealed that their replication time is highly correlated with the distance to centromeres, with early activated ARS preferentially located close to centromeres and depleted towards the telomeres [5]. Since centromeres are clustered at the SPB, this might cause the clustering of early replication origins. Hence, it is not clear if the hypothesis of clustering of early ARS is mechanistic or it is only imposed by the clustering of centromeres in the Rab1-like organization, since those ARS tend to be close to the centromeres. Similarly, some ARS are more clustered during quiescence (far from the centromeric regions) and other are more clustered in exponential

growth (those close to the centromeric regions) but this might also be a consequence of centromere dynamics in quiescent cells.

1.5.4 tRNA genes

Transfer RNA (tRNA) genes are regulated by RNA Pol III and their transcription is initiated by TFIIC protein complex [135]. They are usually bound by SMC proteins, chromatin remodelers and other architectural proteins [136]–[138], showing their importance on chromatin organization. Fluorescent *in situ* hybridization (FISH) studies reported increased contacts between tRNA genes mediated by condensing activity [139]. On the other hand, initial Hi-C studies at low resolution [110] detected a cluster of tRNA genes close to the nucleolus. Hi-C analyses at higher resolution [131] found that groups of tRNA genes are also correlated with the distance to centromere and so divided into two groups: close and far from centromere cluster. This separation into two groups does not change in quiescence, although the group of tRNA genes closer to centromeres tend to have lower interaction frequencies among them. Recently, it was shown that nucleosome positioning, binding of SMC complexes and centromere clustering are affected by the deletion of tRNA genes from an entire chromosome in yeast [140].

1.6 Chromatin modeling

The ability to obtain information about chromatin contacts at the genome-wide scale from Hi-C experiments boosted the development of several physical models aiming to explain its 3D conformation. In general, contact frequencies are assumed to be a proxy for structural proximity. These models arise from the observation that contact frequencies decay at similar rates as observed in polymer physics [141]. They can be divided into two main groups. The first group of ‘bottom-up’ approaches model a hypothesized mechanism of chromatin folding, trying to reproduce the contact probabilities observed from the 3C models. The second group, ‘top-down’ models use the information from the contact frequencies applied as restraints to find possible configurations of the chromatin in 3D, searching to derive possible mechanisms of chromatin folding.

1.6.1 Bottom-up chromatin models

From the observation that TADs are highly related to chromatin modifications, and the clear characterization of chromatin domains in *D. melanogaster* (active, Polycomb-repressed, HP1 repressed or black chromatin), Jost et al. [142] proposed to model chromatin as beads on a string co-polymer (see **Figure 1.8**) imposing larger attraction between beads of the same epigenetic type and repulsion between beads of different type. It was shown that the model could generate chromatin structures that are consistent with Hi-C contact maps in regions of approximately 1.3 Mbp. However, all chromatin types have the same parameters and the model should be refined to account for the different density and compaction of each epigenetic type, as observed from super resolution microscopy [95].

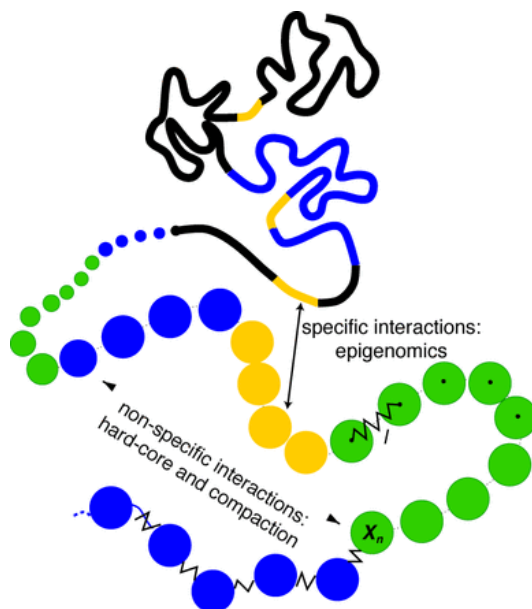


Figure 1.8. Chromatin polymer model as beads on a string array with monomers of 10Kb from epigenomic domains. The color of each bead represents its epigenomic state and is used to define the specific (between beads of the same color) and non-specific (beads of different color) interactions in the model. Taken from [142]

Another approach is the Strings and Binders Switch (SBS) polymer model [143] which introduces the effect of proteins that might act as looping factors, attracting close in space regions containing its recognition motif (**Figure 1.9**). Here, the chromatin is also represented as self-avoiding beads on a string interacting with

binders following a Lennard-Jones attractive potential with a given energy of interaction and concentration of binders. The model can reach a coiled open state at low energy of interaction and concentration, or more closed globular states that can be disordered or organized depending on the magnitude of the energies of interaction and the concentration of binders.

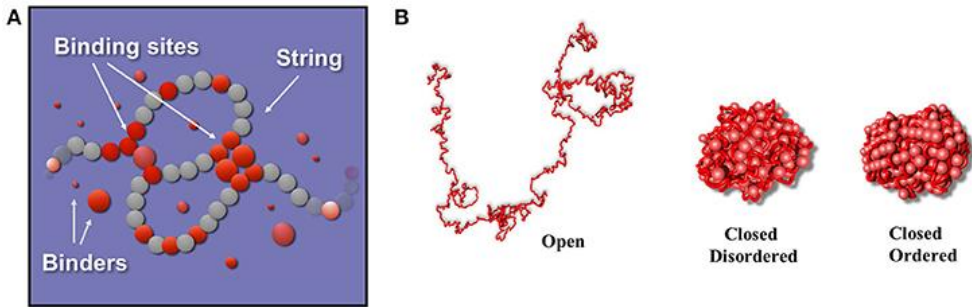


Figure 1.9. Representation of the SBS model of chromatin. (A) Chromatin is organized by binding factors such as transcription factors. (B) Three classes of stable conformations can be obtained: (left) an open Self-Avoiding Walk chain randomly folded, (center) a closed disordered globule state and (right) a closed ordered globule state produced by higher interaction energies or concentrations. Taken from [144]

Other bottom-up approaches include the loop extrusion model [115], [116] that takes into account the effect of cohesin and CTCF to extrude chromatin through the cohesin ring and stop when CTCF bound to DNA is encountered in convergent orientation and models that introduce the effect of supercoiling in 3D structure, for instance by torsional stress produced by RNA polymerases [145].

1.6.2 Top-down chromatin models

Top-down models typically transform the binned contact matrices from 3C-based experiments to spatial restraints that are posteriorly applied to obtain structures that resemble the original contact data to some extent [110], [141]. These models can be validated with other available experimental data such as distances between target loci obtained from FISH, measurements of structure volume, compaction marks obtained from sequencing experiments such as chromatin accessibility or chromatin modifications or information about LADs. In this section, three types of top-down models are summarized, as presented by Polles et al. [146]

The first type of models searches for a **consensus structure** that represents an average conformation from the population data. They minimize the deviations between the distances in the model and those derived from the experimental contact frequencies, assuming that the larger the contact frequency, the shorter the expected distance between each pair of loci. Several methods have been proposed to achieve this minimization, such as scoring function optimization [110], [141], Bayesian likelihood function maximization [147] or generalized linear models [148]. A disadvantage of this strategy is that, since 3C-based data is obtained from a population of cells, the obtained structures are average representations that are not necessarily observed in single cells.

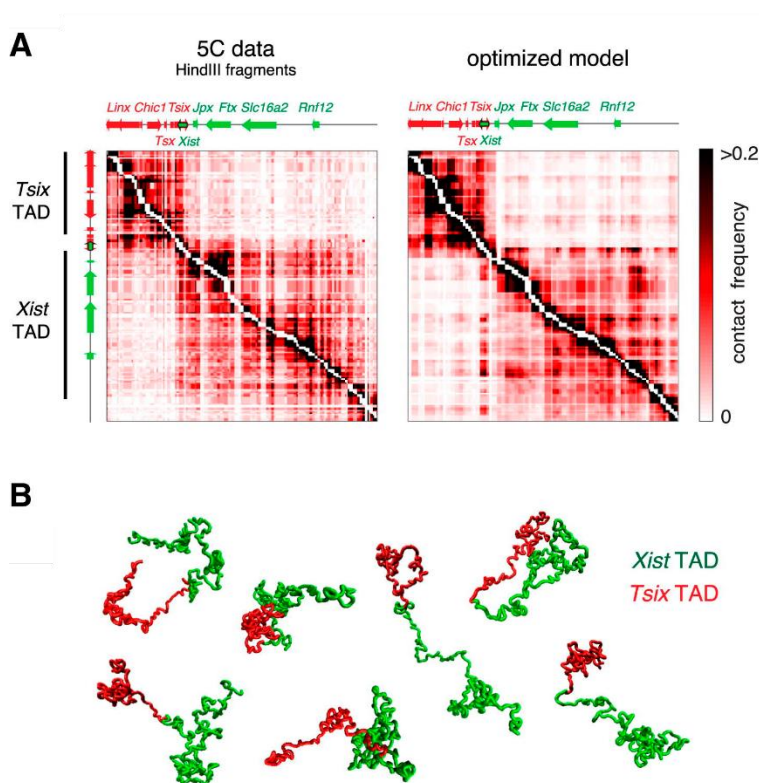


Figure 1.10. Example of results from a top-down chromatin model. (A) Experimental (left panel) and simulated (right panel) contact matrices for a region containing the *Tsix* and *Xist* TADs. (B) Sample conformations from the optimized simulation showing the two TADs in different colors. Adapted from [153]

The second group, **resampling models**, also make a conversion from contact frequencies from the 3C experiment to spatial restraints but obtain an ensemble of structures by defining optimizations with multiple minima or thermodynamic fluctuations. Since the variability in chromatin structure between individual cells is large, as reported from single cell Hi-C and super-resolution microscopy experiments [149], some of these models use only a part of the cells from the contact matrix [150]–[152].

Finally, in population-based **deconvolution methods**, the 3C-based contacts are transformed into single structures that only contain a subset of the original contacts that are conformationally possible obtaining an ensemble of possible configurations (see an example in **Figure 1.10**) [153], [154].

Bibliography for Chapter 1

- [1] B. R. Lajoie, J. Dekker, and N. Kaplan, "The Hitchhiker's guide to Hi-C analysis: Practical guidelines," *Methods*, vol. 72, pp. 65–75, Jan. 2015.
- [2] G. Cavalli and T. Misteli, "Functional implications of genome topology," *Nat. Struct. Mol. Biol.*, vol. 20, no. 3, pp. 290–299, Mar. 2013.
- [3] B. van Steensel and E. E. M. Furlong, "The role of transcription in shaping the spatial organization of the genome," *Nat. Rev. Mol. Cell Biol.*, Mar. 2019.
- [4] M. J. Rowley and V. G. Corces, "Organizational principles of 3D genome architecture," *Nat. Rev. Genet.*, vol. 19, no. 12, pp. 789–800, Dec. 2018.
- [5] L. Lazar-Stefanita *et al.*, "Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle," *EMBO J.*, vol. 36, no. 18, pp. 2684–2697, Sep. 2017.
- [6] E. Vizcaya-Molina, C. C. Klein, F. Serras, and M. Corominas, "Chromatin dynamics in regeneration epithelia: Lessons from *Drosophila* imaginal discs," *Semin. Cell Dev. Biol.*, p. S1084952118301952, May 2019.
- [7] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 Å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, Sep. 1997.
- [8] T. J. Richmond, J. T. Finch, B. Rushton, D. Rhodes, and A. Klug, "Structure of the nucleosome core particle at 7 Å resolution," *Nature*, vol. 311, no. 5986, pp. 532–537, Oct. 1984.
- [9] B. Alaskhar Alhamwe *et al.*, "Histone modifications and their role in epigenetics of atopy and allergic diseases," *Allergy Asthma Clin. Immunol.*, vol. 14, no. 1, Dec. 2018.
- [10] F. Battistini, C. A. Hunter, E. J. Gardiner, and M. J. Packer, "Structural Mechanics of DNA Wrapping in the Nucleosome," *J. Mol. Biol.*, vol. 396, no. 2, pp. 264–279, Feb. 2010.
- [11] R. V. Chereji and D. J. Clark, "Major Determinants of Nucleosome Positioning," *Biophys. J.*, Apr. 2018.
- [12] J. Hu *et al.*, "Dynamic placement of the linker histone H1 associated with nucleosome arrangement and gene transcription in early *Drosophila* embryonic development," *Cell Death Dis.*, vol. 9, no. 7, Jul. 2018.
- [13] W. K. M. Lai and B. F. Pugh, "Understanding nucleosome dynamics and their links to gene expression and DNA replication," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 9, pp. 548–562, May 2017.
- [14] P. G. Giresi, J. Kim, R. M. McDaniel, V. R. Iyer, and J. D. Lieb, "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin," *Genome Res.*, vol. 17, no. 6, pp. 877–885, Jun. 2007.
- [15] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nat. Methods*, vol. 10, no. 12, pp. 1213–1218, Dec. 2013.
- [16] D. E. Schones *et al.*, "Dynamic Regulation of Nucleosome Positioning in the Human Genome," *Cell*, vol. 132, no. 5, pp. 887–898, Mar. 2008.
- [17] O. Flores, O. Deniz, M. Soler-Lopez, and M. Orozco, "Fuzziness and noise in nucleosomal architecture," *Nucleic Acids Res.*, vol. 42, no. 8, pp. 4934–4946, Apr. 2014.
- [18] F. Zhu *et al.*, "The interaction landscape between transcription factors and the nucleosome," *Nature*, Sep. 2018.
- [19] G. Längst and L. Manelyte, "Chromatin Remodelers: From Function to Dysfunction," *Genes*, vol. 6, no. 2, pp. 299–324, Jun. 2015.
- [20] A. Jansen and K. J. Verstrepen, "Nucleosome Positioning in *Saccharomyces cerevisiae*," *Microbiol. Mol. Biol. Rev.*, vol. 75, no. 2, pp. 301–320, Jun. 2011.

- [21] Ö. Deniz, O. Flores, F. Battistini, A. Pérez, M. Soler-López, and M. Orozco, "Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast," *BMC Genomics*, vol. 12, no. 1, Dec. 2011.
- [22] E. Segal *et al.*, "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, pp. 772–778, Aug. 2006.
- [23] A. Pérez *et al.*, "Impact of Methylation on the Physical Properties of DNA," *Biophys. J.*, vol. 102, no. 9, pp. 2140–2148, May 2012.
- [24] V. B. Teif, "Nucleosome positioning: resources and tools online," *Brief. Bioinform.*, vol. 17, no. 5, pp. 745–757, Sep. 2016.
- [25] A. Arneodo, G. Drillon, F. Argoul, and B. Audit, "The Role of Nucleosome Positioning in Genome Function and Evolution," in *Nuclear Architecture and Dynamics*, Elsevier, 2018, pp. 41–79.
- [26] W. Lee *et al.*, "A high-resolution atlas of nucleosome occupancy in yeast," *Nat. Genet.*, vol. 39, no. 10, pp. 1235–1244, Oct. 2007.
- [27] O. J. Rando and K. Ahmad, "Rules and regulation in the primary structure of chromatin," *Curr. Opin. Cell Biol.*, vol. 19, no. 3, pp. 250–256, Jun. 2007.
- [28] C. Jiang and B. F. Pugh, "Nucleosome positioning and gene regulation: advances through genomics," *Nat. Rev. Genet.*, vol. 10, no. 3, pp. 161–172, Mar. 2009.
- [29] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," *Nature*, vol. 474, no. 7352, pp. 516–520, Jun. 2011.
- [30] R. V. Chereji and A. V. Morozov, "Ubiquitous nucleosome crowding in the yeast genome," *Proc. Natl. Acad. Sci.*, vol. 111, no. 14, pp. 5236–5241, Apr. 2014.
- [31] I. M. Nodelman *et al.*, "Interdomain Communication of the Chd1 Chromatin Remodeler across the DNA Gyres of the Nucleosome," *Mol. Cell*, vol. 65, no. 3, pp. 447–459.e6, Feb. 2017.
- [32] I. L. de la Serna, Y. Ohkawa, and A. N. Imbalzano, "Chromatin remodelling in mammalian differentiation: lessons from ATP-dependent remodellers," *Nat. Rev. Genet.*, vol. 7, no. 6, pp. 461–473, Jun. 2006.
- [33] A. Weiner, A. Hughes, M. Yassour, O. J. Rando, and N. Friedman, "High-resolution nucleosome mapping reveals transcription-dependent promoter packaging," *Genome Res.*, vol. 20, no. 1, pp. 90–100, Jan. 2010.
- [34] C. M. Weber, S. Ramachandran, and S. Henikoff, "Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase," *Mol. Cell*, vol. 53, no. 5, pp. 819–830, Mar. 2014.
- [35] K. Struhl and E. Segal, "Determinants of nucleosome positioning," *Nat. Struct. Mol. Biol.*, vol. 20, no. 3, pp. 267–273, Mar. 2013.
- [36] T. Kujirai, H. Ehara, Y. Fujino, M. Shirouzu, S. Sekine, and H. Kurumizaka, "Structural basis of the nucleosome transition during RNA polymerase II passage," *Science*, vol. 362, no. 6414, pp. 595–598, Nov. 2018.
- [37] L. Farnung, S. M. Vos, and P. Cramer, "Structure of transcribing RNA polymerase II-nucleosome complex," *Nat. Commun.*, vol. 9, no. 1, Dec. 2018.
- [38] R. Reja, V. Vinayachandran, S. Ghosh, and B. F. Pugh, "Molecular mechanisms of ribosomal protein gene coregulation," *Genes Dev.*, vol. 29, no. 18, pp. 1942–1954, Sep. 2015.
- [39] A. B. Lantermann, T. Straub, A. Strålfors, G.-C. Yuan, K. Ekwall, and P. Korber, "Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae," *Nat. Struct. Mol. Biol.*, vol. 17, no. 2, pp. 251–257, Feb. 2010.
- [40] V. B. Teif *et al.*, "Genome-wide nucleosome positioning during embryonic stem cell development," *Nat. Struct. Mol. Biol.*, vol. 19, no. 11, pp. 1185–1192, Nov. 2012.

- [41] T. N. Mavrich *et al.*, "Nucleosome organization in the *Drosophila* genome," *Nature*, vol. 453, no. 7193, pp. 358–362, May 2008.
- [42] B. Lai *et al.*, "Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing," *Nature*, vol. 562, no. 7726, pp. 281–285, Oct. 2018.
- [43] S. Baldi, S. Krebs, H. Blum, and P. B. Becker, "Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing," *Nat. Struct. Mol. Biol.*, Aug. 2018.
- [44] C. Vaillant *et al.*, "A novel strategy of transcription regulation by intragenic nucleosome ordering," *Genome Res.*, vol. 20, no. 1, pp. 59–67, Jan. 2010.
- [45] R. V. Chereji, S. Ramachandran, T. D. Bryson, and S. Henikoff, "Precise genome-wide mapping of single nucleosomes and linkers in vivo," *Genome Biol.*, vol. 19, no. 1, Dec. 2018.
- [46] G. Drillon, B. Audit, F. Argoul, and A. Arneodo, "Evidence of selection for an accessible nucleosomal array in human," *BMC Genomics*, vol. 17, no. 1, Dec. 2016.
- [47] M. Weinreich, M. A. Palacios DeBeer, and C. A. Fox, "The activities of eukaryotic replication origins in chromatin," *Biochim. Biophys. Acta BBA - Gene Struct. Expr.*, vol. 1677, no. 1–3, pp. 142–157, Mar. 2004.
- [48] J. J. Wyrick, "Genome-Wide Distribution of ORC and MCM Proteins in *S. cerevisiae*: High-Resolution Mapping of Replication Origins," *Science*, vol. 294, no. 5550, pp. 2357–2360, Dec. 2001.
- [49] M. L. Eaton, K. Galani, S. Kang, S. P. Bell, and D. M. MacAlpine, "Conserved nucleosome positioning defines replication origins," *Genes Dev.*, vol. 24, no. 8, pp. 748–753, Apr. 2010.
- [50] C. Cayrou *et al.*, "The chromatin environment shapes DNA replication origin organization and defines origin classes," *Genome Res.*, vol. 25, no. 12, pp. 1873–1885, Dec. 2015.
- [51] Ö. Deniz, O. Flores, M. Aldea, M. Soler-López, and M. Orozco, "Nucleosome architecture throughout the cell cycle," *Sci. Rep.*, vol. 6, p. 19729, Jan. 2016.
- [52] R. C. Conaway and J. W. Conaway, "The INO80 chromatin remodeling complex in transcription, replication and repair," *Trends Biochem. Sci.*, vol. 34, no. 2, pp. 71–77, Feb. 2009.
- [53] Y. Dalal, T. Furuyama, D. Vermaak, and S. Henikoff, "Structure, dynamics, and evolution of centromeric nucleosomes," *Proc. Natl. Acad. Sci.*, vol. 104, no. 41, pp. 15974–15981, Oct. 2007.
- [54] G. Mizuguchi, H. Xiao, J. Wisniewski, M. M. Smith, and C. Wu, "Nonhistone Scm3 and Histones CenH3-H4 Assemble the Core of Centromere-Specific Nucleosomes," *Cell*, vol. 129, no. 6, pp. 1153–1164, Jun. 2007.
- [55] H. S. Rhee, A. R. Bataille, L. Zhang, and B. F. Pugh, "Subnucleosomal Structures and Nucleosome Asymmetry across a Genome," *Cell*, vol. 159, no. 6, pp. 1377–1388, Dec. 2014.
- [56] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications," *Cell Res.*, vol. 21, no. 3, pp. 381–395, Mar. 2011.
- [57] A. Allahverdi *et al.*, "The effects of histone H4 tail acetylations on cation-induced chromatin folding and self-association," *Nucleic Acids Res.*, vol. 39, no. 5, pp. 1680–1691, Mar. 2011.
- [58] M. Ohno, T. Ando, D. G. Priest, V. Kumar, Y. Yoshida, and Y. Taniguchi, "Sub-nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs," *Cell*, Jan. 2019.
- [59] E. Ceccacci and S. Minucci, "Inhibition of histone deacetylases in cancer therapy: lessons from leukaemia," *Br. J. Cancer*, vol. 114, no. 6, pp. 605–611, Mar. 2016.
- [60] V. M. Weake and J. L. Workman, "Histone Ubiquitination: Triggering Gene Activity," *Mol. Cell*, vol. 29, no. 6, pp. 653–663, Mar. 2008.
- [61] F. Capuano, M. Müllleder, R. Kok, H. J. Blom, and M. Ralser, "Cytosine DNA Methylation Is Found in *Drosophila melanogaster* but Absent in *Saccharomyces cerevisiae*,

- Schizosaccharomyces pombe*, and Other Yeast Species," *Anal. Chem.*, vol. 86, no. 8, pp. 3697–3702, Apr. 2014.
- [62] P. D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery, and M. Orozco, "Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA," *Nucleic Acids Res.*, vol. 42, no. 18, pp. 11304–11320, Oct. 2014.
- [63] Y. Dor and H. Cedar, "Principles of DNA methylation and their implications for biology and medicine," *The Lancet*, vol. 392, no. 10149, pp. 777–786, Sep. 2018.
- [64] Y. Yin *et al.*, "Impact of cytosine methylation on DNA binding specificities of human transcription factors," *Science*, vol. 356, no. 6337, p. eaaj2239, May 2017.
- [65] H. Zhu, G. Wang, and J. Qian, "Transcription factors as readers and effectors of DNA methylation," *Nat. Rev. Genet.*, vol. 17, no. 9, pp. 551–565, Sep. 2016.
- [66] J. F. Kribelbauer *et al.*, "Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes," *Cell Rep.*, vol. 19, no. 11, pp. 2383–2395, Jun. 2017.
- [67] R. Straussman *et al.*, "Developmental programming of CpG island methylation profiles in the human genome," *Nat. Struct. Mol. Biol.*, vol. 16, no. 5, pp. 564–571, May 2009.
- [68] S. Epsztejn-Litman *et al.*, "De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes," *Nat. Struct. Mol. Biol.*, vol. 15, no. 11, pp. 1176–1183, Nov. 2008.
- [69] S. Orlanski *et al.*, "Tissue-specific DNA demethylation is required for proper B-cell differentiation and function," *Proc. Natl. Acad. Sci.*, vol. 113, no. 18, pp. 5018–5023, May 2016.
- [70] H. Heyn *et al.*, "Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient," *Epigenetics*, vol. 7, no. 6, pp. 542–550, Jun. 2012.
- [71] J. Winkelmann *et al.*, "Mutations in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy," *Hum. Mol. Genet.*, vol. 21, no. 10, pp. 2205–2210, May 2012.
- [72] C. Holz-Schietinger, D. M. Matje, and N. O. Reich, "Mutations in DNA Methyltransferase (DNMT3A) Observed in Acute Myeloid Leukemia Patients Disrupt Processive Methylation," *J. Biol. Chem.*, vol. 287, no. 37, pp. 30941–30951, Sep. 2012.
- [73] M. Kulis *et al.*, "Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia," *Nat. Genet.*, vol. 44, no. 11, pp. 1236–1242, Oct. 2012.
- [74] E. Vidal *et al.*, "A DNA methylation map of human cancer at single base-pair resolution," *Oncogene*, vol. 36, no. 40, pp. 5648–5657, Oct. 2017.
- [75] D. Aran and A. Hellman, "DNA Methylation of Transcriptional Enhancers and Cancer Predisposition," *Cell*, vol. 154, no. 1, pp. 11–13, Jul. 2013.
- [76] M. Klutstein, D. Nejman, R. Greenfield, and H. Cedar, "DNA Methylation in Cancer and Aging," *Cancer Res.*, vol. 76, no. 12, pp. 3446–3450, Jun. 2016.
- [77] S. Ecker, V. Pancaldi, A. Valencia, S. Beck, and D. S. Paul, "Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity," *BioEssays*, vol. 40, no. 2, p. 1700148, Feb. 2018.
- [78] A. Portela, J. Liz, V. Nogales, F. Setién, A. Villanueva, and M. Esteller, "DNA methylation determines nucleosome occupancy in the 5'-CpG islands of tumor suppressor genes," *Oncogene*, vol. 32, no. 47, pp. 5421–5428, Nov. 2013.
- [79] T. T. M. Ngo *et al.*, "Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability," *Nat. Commun.*, vol. 7, p. 10813, Feb. 2016.
- [80] C. K. Collings and J. N. Anderson, "Links between DNA methylation and nucleosome occupancy in the human genome," *Epigenetics Chromatin*, vol. 10, no. 1, Dec. 2017.
- [81] R. K. Chodavarapu *et al.*, "Relationship between nucleosome positioning and DNA methylation," *Nature*, vol. 466, no. 7304, pp. 388–392, Jul. 2010.
- [82] M. Felle, H. Hoffmeister, J. Rothammer, A. Fuchs, J. H. Exler, and G. Langst, "Nucleosomes protect DNA from DNA methylation in vivo and in vitro," *Nucleic Acids Res.*, vol. 39, no. 16, pp. 6956–6969, Sep. 2011.

- [83] J. T. Huff and D. Zilberman, "Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes," *Cell*, vol. 156, no. 6, pp. 1286–1297, Mar. 2014.
- [84] T. K. Kelly, Y. Liu, F. D. Lay, G. Liang, B. P. Berman, and P. A. Jones, "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules," *Genome Res.*, vol. 22, no. 12, pp. 2497–2506, Dec. 2012.
- [85] D. A. Beshnova, A. G. Cherstvy, Y. Vainshtein, and V. B. Teif, "Regulation of the Nucleosome Repeat Length In Vivo by the DNA Sequence, Protein Concentrations and Long-Range Interactions," *PLoS Comput. Biol.*, vol. 10, no. 7, p. e1003698, Jul. 2014.
- [86] A. L. Olins and D. E. Olins, "Spheroid chromatin units (v bodies)," *Science*, vol. 183, no. 4122, pp. 330–332, Jan. 1974.
- [87] R. D. Kornberg, "Chromatin Structure: A Repeating Unit of Histones and DNA," *Science*, vol. 184, no. 4139, pp. 868–871, May 1974.
- [88] M. Ohno, D. G. Priest, and Y. Taniguchi, "Nucleosome-level 3D organization of the genome," *Biochem. Soc. Trans.*, vol. 46, no. 3, pp. 491–501, Jun. 2018.
- [89] J. T. Finch and A. Klug, "Solenoidal model for superstructure in chromatin.," *Proc. Natl. Acad. Sci.*, vol. 73, no. 6, pp. 1897–1901, Jun. 1976.
- [90] C. Woodcock, "The higher-order structure of chromatin: evidence for a helical ribbon arrangement," *J. Cell Biol.*, vol. 99, no. 1, pp. 42–52, Jul. 1984.
- [91] S. P. Williams, B. D. Athey, L. J. Muglia, R. S. Schappe, A. H. Gough, and J. P. Langmore, "Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length," *Biophys. J.*, vol. 49, no. 1, pp. 233–248, Jan. 1986.
- [92] B. Dorigo, "Nucleosome Arrays Reveal the Two-Start Organization of the Chromatin Fiber," *Science*, vol. 306, no. 5701, pp. 1571–1573, Nov. 2004.
- [93] E. Fussner, R. W. Ching, and D. P. Bazett-Jones, "Living without 30nm chromatin fibers," *Trends Biochem. Sci.*, vol. 36, no. 1, pp. 1–6, Jan. 2011.
- [94] K. Maeshima, S. Hihara, and M. Eltsov, "Chromatin structure: does the 30-nm fibre exist in vivo?," *Curr. Opin. Cell Biol.*, vol. 22, no. 3, pp. 291–297, Jun. 2010.
- [95] A. N. Boettiger *et al.*, "Super-resolution imaging reveals distinct chromatin folding for different epigenetic states," *Nature*, vol. 529, no. 7586, pp. 418–422, Jan. 2016.
- [96] M. A. Ricci, C. Manzo, M. F. García-Parajo, M. Lakadamyali, and M. P. Cosma, "Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo," *Cell*, vol. 160, no. 6, pp. 1145–1158, Mar. 2015.
- [97] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, "Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C," *Cell*, vol. 162, no. 1, pp. 108–119, Jul. 2015.
- [98] T.-H. S. Hsieh, G. Fudenberg, A. Goloborodko, and O. J. Rando, "Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome," *Nat. Methods*, vol. 13, no. 12, pp. 1009–1011, Oct. 2016.
- [99] V. I. Risca, S. K. Denny, A. F. Straight, and W. J. Greenleaf, "Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping," *Nature*, vol. 541, no. 7636, pp. 237–241, Jan. 2017.
- [100] E. Lieberman-Aiden *et al.*, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009.
- [101] J. H. Gibcus and J. Dekker, "The Hierarchy of the 3D Genome," *Mol. Cell*, vol. 49, no. 5, pp. 773–782, Mar. 2013.
- [102] J. Fraser *et al.*, "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation," *Mol. Syst. Biol.*, vol. 11, no. 12, pp. 852–852, Dec. 2015.
- [103] T. Cremer and C. Cremer, "Chromosome territories, nuclear architecture and gene regulation in mammalian cells," *Nat. Rev. Genet.*, vol. 2, no. 4, pp. 292–301, Apr. 2001.

- [104] B. van Steensel and A. S. Belmont, "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression," *Cell*, vol. 169, no. 5, pp. 780–791, May 2017.
- [105] R. Pieschel, F. Coraggio, and P. Meister, "From single genes to entire genomes: the search for a function of nuclear organization," *Development*, vol. 143, no. 6, pp. 910–923, Mar. 2016.
- [106] E. P. Nora *et al.*, "Spatial partitioning of the regulatory landscape of the X-inactivation centre," *Nature*, vol. 485, no. 7398, pp. 381–385, Apr. 2012.
- [107] S. M. Tan-Wong *et al.*, "Gene Loops Enhance Transcriptional Directionality," *Science*, vol. 338, no. 6107, pp. 671–675, Nov. 2012.
- [108] T. Sexton *et al.*, "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome," *Cell*, vol. 148, no. 3, pp. 458–472, Feb. 2012.
- [109] R. A. Beagrie *et al.*, "Complex multi-enhancer contacts captured by genome architecture mapping," *Nature*, vol. 543, no. 7646, pp. 519–524, Mar. 2017.
- [110] Z. Duan *et al.*, "A three-dimensional model of the yeast genome," *Nature*, vol. 465, no. 7296, pp. 363–367, May 2010.
- [111] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, Apr. 2012.
- [112] J.-P. Fortin and K. D. Hansen, "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data," *Genome Biol.*, vol. 16, no. 1, Dec. 2015.
- [113] S. Nothjunge *et al.*, "DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes," *Nat. Commun.*, vol. 8, no. 1, Dec. 2017.
- [114] S. S. P. Rao *et al.*, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014.
- [115] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, "Formation of Chromosomal Domains by Loop Extrusion," *Cell Rep.*, vol. 15, no. 9, pp. 2038–2049, May 2016.
- [116] A. L. Sanborn *et al.*, "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes," *Proc. Natl. Acad. Sci.*, vol. 112, no. 47, pp. E6456–E6465, Nov. 2015.
- [117] S. S. P. Rao *et al.*, "Cohesin Loss Eliminates All Loop Domains," *Cell*, vol. 171, no. 2, pp. 305–320.e24, Oct. 2017.
- [118] D. G. Lupiáñez, M. Spielmann, and S. Mundlos, "Breaking TADs: How Alterations of Chromatin Domains Result in Disease," *Trends Genet.*, vol. 32, no. 4, pp. 225–237, Apr. 2016.
- [119] F. Ramírez *et al.*, "High-resolution TADs reveal DNA sequences underlying genome organization in flies," *Nat. Commun.*, vol. 9, no. 1, Dec. 2018.
- [120] A.-L. Valton and J. Dekker, "TAD disruption as oncogenic driver," *Curr. Opin. Genet. Dev.*, vol. 36, pp. 34–40, Feb. 2016.
- [121] T. Sexton and G. Cavalli, "The Role of Chromosome Domains in Shaping the Functional Genome," *Cell*, vol. 160, no. 6, pp. 1049–1059, Mar. 2015.
- [122] M. J. Rowley *et al.*, "Evolutionarily Conserved Principles Predict 3D Chromatin Organization," *Mol. Cell*, vol. 67, no. 5, pp. 837–852.e7, Sep. 2017.
- [123] R. Wang, J. Mozziconacci, A. Bancaud, and O. Gadal, "Principles of chromatin organization in yeast: relevance of polymer models to describe nuclear organization and dynamics," *Curr. Opin. Cell Biol.*, vol. 34, pp. 54–60, Jun. 2015.
- [124] K. Bystricky, T. Laroche, G. van Houwe, M. Blaszczyk, and S. M. Gasser, "Chromosome looping in yeast: telomere pairing and coordinated movement reflect anchoring efficiency and territorial organization," *J. Cell Biol.*, vol. 168, no. 3, pp. 375–387, Jan. 2005.
- [125] A. Taddei and S. M. Gasser, "Structure and Function in the Budding Yeast Nucleus," *Genetics*, vol. 192, no. 1, pp. 107–129, Sep. 2012.

- [126] P. Therizols, T. Duong, B. Dujon, C. Zimmer, and E. Fabre, "Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres," *Proc. Natl. Acad. Sci.*, vol. 107, no. 5, pp. 2025–2030, Feb. 2010.
- [127] J.-M. Belton *et al.*, "The Conformation of Yeast Chromosome III Is Mating Type Dependent and Controlled by the Recombination Enhancer," *Cell Rep.*, vol. 13, no. 9, pp. 1855–1867, Dec. 2015.
- [128] S. A. Schalbetter *et al.*, "SMC complexes differentially compact mitotic chromosomes according to genomic context," *Nat. Cell Biol.*, vol. 19, no. 9, pp. 1071–1080, Aug. 2017.
- [129] Y. Kakui and F. Uhlmann, "SMC complexes orchestrate the mitotic chromatin interaction landscape," *Curr. Genet.*, vol. 64, no. 2, pp. 335–339, Apr. 2018.
- [130] D. Wang, A. Mansisidor, G. Prabhakar, and A. Hochwagen, "Condensin and Hmo1 Mediate a Starvation-Induced Transcriptional Position Effect within the Ribosomal DNA Array," *Cell Rep.*, vol. 14, no. 5, pp. 1010–1017, Feb. 2016.
- [131] M. T. Rutledge, M. Russo, J.-M. Belton, J. Dekker, and J. R. Broach, "The yeast genome undergoes significant topological reorganization in quiescence," *Nucleic Acids Res.*, vol. 43, no. 17, pp. 8299–8313, Sep. 2015.
- [132] C. K. Tsang, H. Li, and X. S. Zheng, "Nutrient starvation promotes condensin loading to maintain rDNA stability," *EMBO J.*, vol. 26, no. 2, pp. 448–458, Jan. 2007.
- [133] M. Guidi *et al.*, "Spatial reorganization of telomeres in long-lived quiescent cells," *Genome Biol.*, vol. 16, no. 1, Dec. 2015.
- [134] N. Saner *et al.*, "Stochastic association of neighboring replicons creates replication factories in budding yeast," *J. Cell Biol.*, vol. 202, no. 7, pp. 1001–1012, Sep. 2013.
- [135] M. Cie la, E. Skowronek, and M. Boguta, "Function of TFIIC, RNA polymerase III initiation factor, in activation and repression of tRNA gene transcription," *Nucleic Acids Res.*, vol. 46, no. 18, pp. 9444–9455, Oct. 2018.
- [136] J. Huang and B. C. Laurent, "A Role for the RSC chromatin remodeler in regulating cohesion of sister chromatid arms," *Cell Cycle Georget. Tex.*, vol. 3, no. 8, pp. 973–975, Aug. 2004.
- [137] C. D'Ambrosio *et al.*, "Identification of cis-acting sites for condensin loading onto budding yeast chromosomes," *Genes Dev.*, vol. 22, no. 16, pp. 2215–2227, Aug. 2008.
- [138] E. F. Glynn *et al.*, "Genome-Wide Mapping of the Cohesin Complex in the Yeast *Saccharomyces cerevisiae*," *PLoS Biol.*, vol. 2, no. 9, p. e259, Jul. 2004.
- [139] M. Thompson, R. A. Haeusler, P. D. Good, and D. R. Engelke, "Nucleolar Clustering of Dispersed tRNA Genes," *Science*, vol. 302, no. 5649, pp. 1399–1401, Nov. 2003.
- [140] O. Hamdani *et al.*, "tRNA Genes Affect Chromosome Structure and Function via Local Effects," *Mol. Cell. Biol.*, vol. 39, no. 8, pp. e00432-18, /mcb/39/8/MCB.00432-18.atom, Feb. 2019.
- [141] N. Varoquaux, F. Ay, W. S. Noble, and J.-P. Vert, "A statistical approach for inferring the 3D structure of the genome," *Bioinformatics*, vol. 30, no. 12, pp. i26–i33, Jun. 2014.
- [142] D. Jost, P. Carrivain, G. Cavalli, and C. Vaillant, "Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains," *Nucleic Acids Res.*, vol. 42, no. 15, pp. 9553–9561, Sep. 2014.
- [143] S. Bianco *et al.*, "Polymer physics predicts the effects of structural variants on chromatin architecture," *Nat. Genet.*, Apr. 2018.
- [144] A. M. Chiariello *et al.*, "A Polymer Physics Investigation of the Architecture of the Murine Orthologue of the 7q11.23 Human Locus," *Front. Neurosci.*, vol. 11, p. 559, Oct. 2017.
- [145] D. Racko, F. Benedetti, J. Dorier, and A. Stasiak, "Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes," *Nucleic Acids Res.*, vol. 46, no. 4, pp. 1648–1660, Feb. 2018.
- [146] G. Polles, N. Hua, A. Yildirim, and F. Alber, "Genome Structure Calculation through Comprehensive Data Integration," in *Modeling the 3D Conformation of Genomes*, 1st ed.,

- G. Tiana and L. Giorgetti, Eds. Boca Raton : Taylor & Francis, 2018. | Series: Series in computational biophysics ; 4: CRC Press, 2019, pp. 253–284.
- [147] M. Hu *et al.*, “Bayesian Inference of Spatial Organizations of Chromosomes,” *PLoS Comput. Biol.*, vol. 9, no. 1, p. e1002893, Jan. 2013.
- [148] Z. Zhang, G. Li, K.-C. Toh, and W.-K. Sung, “3D Chromosome Modeling with Semi-Definite Programming and Hi-C Data,” *J. Comput. Biol.*, vol. 20, no. 11, pp. 831–846, Nov. 2013.
- [149] T. Nagano *et al.*, “Cell-cycle dynamics of chromosomal organization at single-cell resolution,” *Nature*, vol. 547, no. 7661, pp. 61–67, Jul. 2017.
- [150] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. J. Filion, and M. A. Marti-Renom, “Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors,” *PLOS Comput. Biol.*, vol. 13, no. 7, p. e1005665, Jul. 2017.
- [151] M. Di Stefano, J. Paulsen, T. G. Lien, E. Hovig, and C. Micheletti, “Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization,” *Sci. Rep.*, vol. 6, p. 35985, Oct. 2016.
- [152] J. Paulsen *et al.*, “Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts,” *Genome Biol.*, vol. 18, no. 1, Dec. 2017.
- [153] L. Giorgetti *et al.*, “Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription,” *Cell*, vol. 157, no. 4, pp. 950–963, May 2014.
- [154] H. Tjong *et al.*, “Population-based 3D genome structure analysis reveals driving forces in spatial genome organization,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 12, pp. E1663–E1672, Mar. 2016.

Objectives

The main objective of this thesis is to study the structure and organization of the DNA fiber at different levels of detail, from local sequence specific properties to global 3D structure within the nucleus. For this purpose, the following specific objectives are proposed and grouped in three categories:

1. DNA sequence dependent properties
 - To characterize the genome wide distribution and function of highly flexible DNA sequences.
 - To assess mechanisms for protein-DNA recognitions defining statistical tests for the detection of significant differences in the physical DNA descriptors between experimental protein-bound and naked DNA structure from molecular dynamics simulations.
 - To predict nucleosome organization profiles using machine learning methods based on the deformation energy of the DNA, transcription factor affinity and periodicity of the nucleosome signal.
2. Tools to study nucleosome positioning *in vivo*
 - To develop an algorithm for comparing nucleosome positioning profiles between two cell populations.
 - To integrate different tools for the analysis of nucleosome organization into a pipeline available through different distribution models (web-

servers, containerized distributions) facilitating the analysis of results in the context of other genomic information.

3. Effect of DNA methylation on chromatin structure

- To analyze the effect of DNA methylation on nucleosome positioning *in vivo*, applying the proposed algorithm for comparison of nucleosome profiles.
- To study the chromatin changes at whole-genome 3D structure level applying statistical methods for the detection of differential interacting regions in Hi-C data.
- To develop a coarse-grained 3D model of the chromatin based on restraints obtained from Hi-C contact matrices for further analysis of the structural changes produced by the DNA methylation.

Chapter 2 . Methods

This chapter presents a summary of the main methods used in this thesis for the analysis of the physical properties of the DNA, chromatin structure and gene expression. Further details about the usage of each method and experimental details can be found in the *Results* section, and in the corresponding papers.

2.1 DNA physical properties

The DNA is an oligomer of nucleotides, forming a double helix where the base pairs are joined by hydrogen bonds. Every base pair step (two consecutive base pairs along the DNA sequence) can be described in the helical space by three translational (shift (f), slide (l), rise (s)) and three rotational (tilt (t), roll (r), twist (w)) movements (see **Figure 2.1**). These physical and geometrical descriptors were derived from molecular dynamics (MD) simulations and were used to study the sequence-dependent DNA equilibrium conformation and deformability at the base pair step level, and to evaluate protein-DNA complex formation energy.

The equilibrium values and stiffness constants for each individual base pair step were taken from MD simulations that cover all the unique base pair steps in all the possible tetranucleotide environments from microsecond-long parmbsc1 simulations [1]. For this, the DNA geometries extracted from the MD simulations were projected into a helical reference system. By collecting the values of these helical parameters, a covariance matrix (C) for each unique base pair step was obtained as follows:

$$C = \frac{\sum_{k=1}^n (x_{ik} - \mu_i)(x_{jk} - \mu_j)}{N - 1} \quad (2.1)$$

where x_{ik} and x_{jk} are the base pair step parameter values at frame k , i and j are one of the six movements (shift, slide, rise, tilt, roll, twist), μ_i and μ_j are their corresponding means, and n is the total number of frames analyzed.

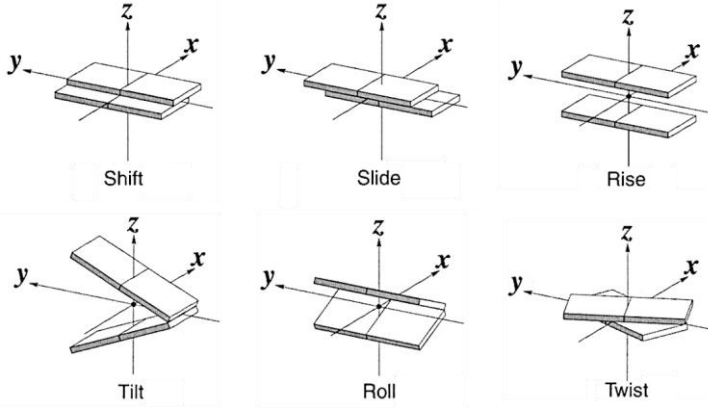


Figure 2.1. Base pair step helical parameters representation. Adapted from [2]

The inverse of this covariance matrix was used to obtain the elastic force constants that represent the energetic cost of the deformation of the DNA molecule along the helical coordinates (eq. 2.2).

$$\theta = k_B T C^{-1} = \begin{bmatrix} k_w & k_{wr} & k_{wt} & k_{ws} & k_{wl} & k_{wf} \\ k_{wr} & k_r & k_{rt} & k_{rs} & k_{rl} & k_{rf} \\ k_{wt} & k_{rt} & k_t & k_{ts} & k_{tl} & k_{tf} \\ k_{ws} & k_{rs} & k_{ts} & k_s & k_{sl} & k_{sf} \\ k_{wl} & k_{rl} & k_{tl} & k_{sl} & k_l & k_{lf} \\ k_{wf} & k_{rf} & k_{tf} & k_{sf} & k_{lf} & k_f \end{bmatrix} \quad (2.2)$$

where k_B is the Boltzmann constant and T is the absolute temperature.

The intrinsic properties of naked DNA that favor protein binding, for instance transcription factor binding at promoters or nucleosomal DNA binding around histones, can be characterized using the stiffness matrix. The energy associated to the deformation of a given base pair step j was computed using a harmonic approximation, given by

$$E_j = \frac{1}{2} \sum_{s=1}^6 \sum_{t=1}^6 k_{st}^j \Delta X_s^j \Delta X_t^j \quad (2.3)$$

where k_{st}^j are the stiffness constants associated with the displacements with respect to the equilibrium values (eq. 2.2) and ΔX_s^j and ΔX_t^j are the differences between the equilibrium values and the protein-bound DNA conformation for the 6 base pair step helical parameters [3], [4].

Finally, the deformation energy associated to the DNA transition from the naked conformation to the protein-bound conformation was calculated in the harmonic regime using (eq. 2.4):

$$Def. Energy = \frac{\sum_{j=1}^m E_j}{m} \quad (2.4)$$

where j stands for each of the m base pair steps of the DNA stretches ($m = 147$ in the case of the deformation energy to wrap around the histones) and E_j is the elastic energy required at each base pair step (eq. 2.3).

2.2 Transcription factor binding

Transcription factors (TFs) are proteins that can recognize and bind to specific sequence motifs to control the expression of genes. As explained above, the protein binding propensities to given DNA sequences can be theoretically studied from the deformation energy associated to the binding process. Additionally, transcription factor binding site (TFBS) affinity can be studied from experiments, such as ChIP-seq (see Section 2.6 for a description of the technique), identifying the position of specific proteins along the genome. This type of analyses have been extensively studied for many TFs in different organisms, and summarized in several databases such as TRANSFAC [5] or JASPAR [6]. We used the binding affinities from JASPAR, given as matrices for each TF containing the frequency of every base (A, C, T or G) in every position of the sequences where the TF is bound. These frequencies were transformed into position weight matrices (PWM) containing normalized scores (in log-scale) [7]. Then, the binding affinity of a given TF to a DNA sequence was estimated adding the corresponding nucleotide values in the PWM. Binding site affinities from every

PWM in JASPAR database were computed for the yeast genome, using R/Bioconductor Biostrings [8] library with default parameters. Finally, a global score of TFBS density was computed pooling the affinities from all TFs in the database at every genome position.

2.3 Nucleosome positioning

The physical properties of DNA can be used to theoretically estimate the propensity of a genomic sequence to form a nucleosome (see Section 2.1). However, other *trans*-factors are important for the *in vivo* positioning of nucleosomes such as the effect of nucleosome remodeler proteins and the local competition with transcription machinery. Hence, methods to determine nucleosome positions are required to be able to study differences in chromatin organization between experimental conditions or cell types.

2.3.1 Studying nucleosome positioning *in vivo*

The most widely used techniques to experimentally map nucleosome positions typically treat a group of cells (in the range 10^6 - 10^9) with enzymes acting on nucleosome-free DNA, after cross-linking with formaldehyde (see **Figure 2.2**). Micrococcal nuclease (MNase) is used to degrade linker DNA preserving preferentially the DNA segments wrapped in the nucleosomes [9], [10]. To obtain the DNA fragments for posterior analysis, the cross-linking is reversed, and the proteins and RNA are digested. Finally, the segments of DNA are sequenced.

Although MNase can be affected by the enzyme concentration and sequence-preference biases limiting the detection of the so called “fragile” nucleosomes, it is the most widely used technique to detect nucleosome positioning for its versatility and accuracy [11]. Chemical cleavage methods have been proposed for circumventing the limitations from MNase-seq, but it requires to do genetic engineering replacing the endogenous histone H4 (or H3 (23)) by a mutated version, therefore restricting its use [12], [13](24–27). Moreover, it has been shown that the

MNase sequence bias can be corrected using digested naked DNA as baseline [14], [15](20, 21), obtaining more pronounced nucleosome coverage peaks.

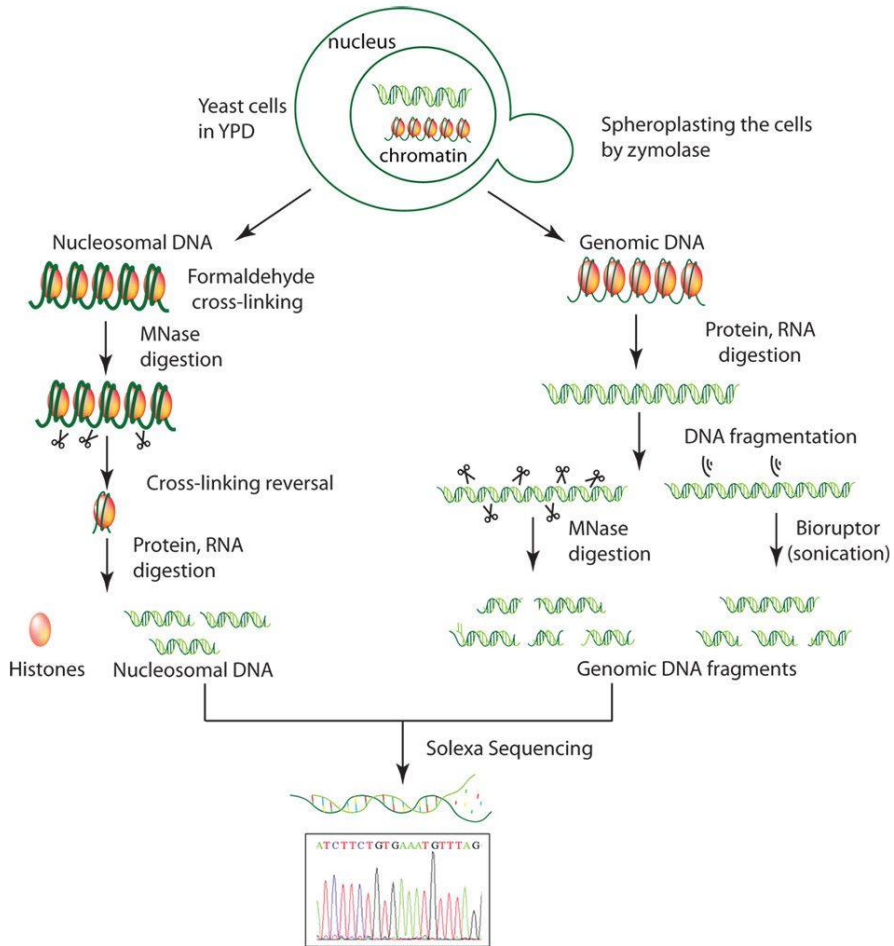


Figure 2.2. MNase-seq experimental procedure. Adapted from [14]

Additionally, the level of MNase digestion should be optimized for each sample to obtain approximately 80% of mono-nucleosomes, using different MNase digestion times with a small amount of semi-intact cells from every batch preparation. The percentage of mono-nucleosomal DNA fragments is examined in agarose gels and the integrity and size distribution of digested fragments are determined using the microfluidics-based platform Bioanalyzer (Agilent) prior to sample preparations and sequencing. The samples are prepared for whole genome sequencing, following

corresponding standard protocols from sequencer manufacturers and the libraries are paired-end sequenced.

The depth of sequencing required for obtaining good quality nucleosome maps should be very high [16], which implies a high sequencing cost. Therefore, a modification of the experimental procedure, Capture-MNase-seq, can be useful to reduce the cost by focusing on target regions in large genomes. Probes designed to hybridize to the objective sequences are added to capture and enrich preferentially the fragments corresponding to the target region prior to sequencing.

2.3.2 Mapping and noise filtering of the MNase signal

The reference genomes corresponding to the samples used along this work were obtained from UCSC: sacCer3 (Apr. 2011, S288C) for yeast samples and hg19 (Feb. 2009, GRCh37) for human cells. Sequenced reads stored in FASTQ files are mapped to the corresponding reference genome using Bowtie [17] aligner, allowing up to two mismatches and an insert length of 500 bp. Reads aligned to multiple regions in the genome are suppressed. The obtained BAM file contains the positions of the reads mapped to the genome and their quality of the alignment. BAM files can be visualized as continuous tracks containing the depth of the coverage at every base pair across the genome (see **Figure 2.3** A and B). Quality control is performed with htSeqTools R/Bioconductor package to remove PCR over-amplification artifacts [18].

2.3.3 Nucleosome calling with nucleR

Mapped fragments that pass the quality control filters are then processed with R/Bioconductor package nucleR [19] as follows (steps to run nucleR can be found at https://github.com/nucleosome-dynamics/nucleosome_dynamics/blob/master/bin/nucleR.R):

- i. Fragments wider than 170 are discarded to keep only those corresponding to mono-nucleosomes.

- ii. Fragments are trimmed to 50bp maintaining the original center to remove noise from MNase digestion variability among cells and regions in the nucleosome coverage profile (see **Figure 2.3 C**).
- iii. The nucleosome coverage per base pair is computed genome-wide and transformed to reads per million mapped.
- iv. Noise is filtered through Fast Fourier Transform, keeping 1% or 2% of the principal components in human and yeast experiments, respectively (see **Figure 2.3 D**).
- v. Finally, peak calling is performed using the parameters: peak width 147 bp, peak detection threshold 35%, maximum overlap 80 bp, dyad length 50 bp. Nucleosome calls are considered well-positioned (W) when nucleR peak width score and height score are higher than 0.6 and 0.4, respectively, and as fuzzy (F) otherwise (see **Figure 2.3 E**).

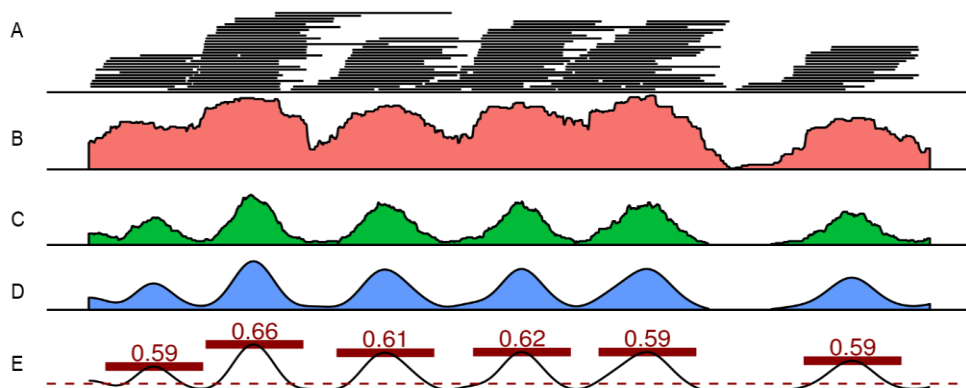


Figure 2.3. Nucleosome positioning from MNase-seq data with nucleR. (A) Reads are mapped to the reference genome. (B) The coverage of nucleosomal reads per base pair is noisy and must be processed further. (C) Reads are trimmed around their center to remove experimental noise, and the coverage is re-computed. (D) Signal is smoothed with Fast Fourier Transform. (E) Peaks are identified from the local maxima and scored according to their height and width. Adapted from [20]

2.4 Chromatin 3D structure

In this section, I describe some methods that have been developed to study 3D chromatin organization. First, I describe three experimental techniques that capture interactions between genomic loci at different resolution and scale. Then, I present the computational processing of the data to remove experimental artifacts, obtain

quantification of the interactions between pairs of regions and find significant differential interacting regions among experimental conditions. Finally, I present some tools that were used in the different projects to visualize the interactions and put them into the corresponding genomic context.

2.4.1 Chromosome conformation capture

Chromosome Conformation Capture (3C) allows us to quantify the frequency of interaction between pairs of loci by crosslinking chromatin, DNA fragmentation and ligation of ends in spatial proximity. Further developments of 3C were proposed to investigate the 3D conformation of chromatin in a population of cells at larger scale. Circular Chromosome Conformation Capture (4C) aims to detect all regions that interact with a single locus of interest [21], [22], Chromosome Conformation Capture Carbon Copy (5C) detects contacts between fragments located within a chromosomal domain of size up to several Mbp [23], Hi-C interrogates genome-wide contacts of all vs. all regions of the genome [24] and finally Capture Hi-C restricts the analysis to contacts between regions targeted by designed probes [25], [26]. Micro-C is another 3C-based technique to quantify genome-wide contacts, but at the nucleosome level resolution due to the DNA cleavage with MNase instead of a restriction enzyme [27].

2.4.1.1 Hi-C

The Hi-C experimental protocol was originally proposed in [24]. In this technique (summarized in **Figure 2.4**), the chromatin is cross-linked with formaldehyde to obtain a static view of its conformation. Then, it is fragmented with a restriction enzyme that recognizes a target DNA motif. The selection of the restriction enzyme will determine the maximum resolution attainable in a given experiment. Four base pair cutters will cut more frequently producing shorter fragments and larger resolution than six base pair cutters.

Next, the ends are filled with nucleotides marked with biotin to facilitate posterior selection of the actual interactions. The fragments in proximity are ligated producing chimeric molecules formed by the two regions that were previously cross-linked. Then, the DNA is purified and sheared by sonication producing fragments of size

appropriate for next generation sequencing. Sonication is not specific and, apart from the chimeric fragments formed by spatial proximity, it produces fragments corresponding to only one region in the genome. Those uninformative fragments can be discarded, since ligation products were previously marked with biotin and can be pulled down and paired-end sequenced.

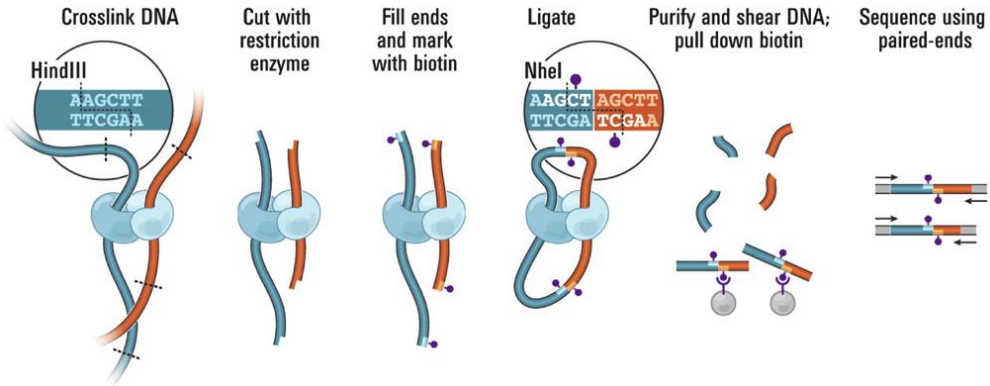


Figure 2.4. Overview of the steps to perform a Hi-C experiment. Adapted from [24]

2.4.1.2 Capture Hi-C

Capture Hi-C is a variation of the Hi-C protocol to enrich the contacts in specific genomic regions. A pool of primers is designed to selectively purify a set of regions and can be used to enrich Hi-C ligation product libraries.

2.4.1.3 Micro-C

Micro-C is another 3C-derived technique aiming to target nucleosome-nucleosome interactions. For this purpose, DNA is fragmented with MNase instead of using a restriction enzyme. Linker DNA is then preferentially cleaved, and the obtained chimeric reads contain sequences corresponding to two nucleosomes that are in spatial proximity.

2.4.2 Quantification of contact frequencies

Several computational algorithms are available for processing the paired-end sequences obtained from 3C experiments (reviewed in [28], [29]). In this work we used TADbit [30], a python library designed for mapping of the paired reads,

filtering and quantification of the obtained contacts and further analyses on the interaction matrices. Below, the steps to process the sequenced reads are explained.

2.4.2.1 Quality control

The quality control was performed also by TADbit, using an algorithm that is based on FastQC program [31] and which checks the PHRED score [32] in the input FASTQ files and the number of “N” positions as a function of the sequence position in the reads. Additionally, TADbit generates plots for the number of undigested sites, dangling ends and re-ligated sites as a function of the nucleotide position in the reads.

2.4.2.2 Mapping

Reads are mapped to the reference genome using GEM mapper [33]. The mapping algorithm must consider that the ligation junction might be contained in any part of the read; therefore, the full length of each read side might not be successfully mapped to the genome. Two mapping strategies accounting for this problem are available in TADbit (see **Figure 2.5**):

- Iterative mapping: The first 25 bp at the 5' end of each read are mapped to the genome. If this sequence is not uniquely mapped, then it is extended 5 bp more and a second attempt to uniquely map it is performed. The process is iterated adding 5 bp each step until either a unique match is found, or the full length of the read is achieved.
- Fragment-based mapping: contrary to iterative mapping, the full length of the fragment is mapped first. Those fragments that fail to be mapped to the genome are split searching for the ligation site, which is known from the motif targeted by the restriction enzyme used in the experiment. It should be noticed that this strategy is not applied to Micro-C experiments, since the MNase digestion is not specific to a given sequence.

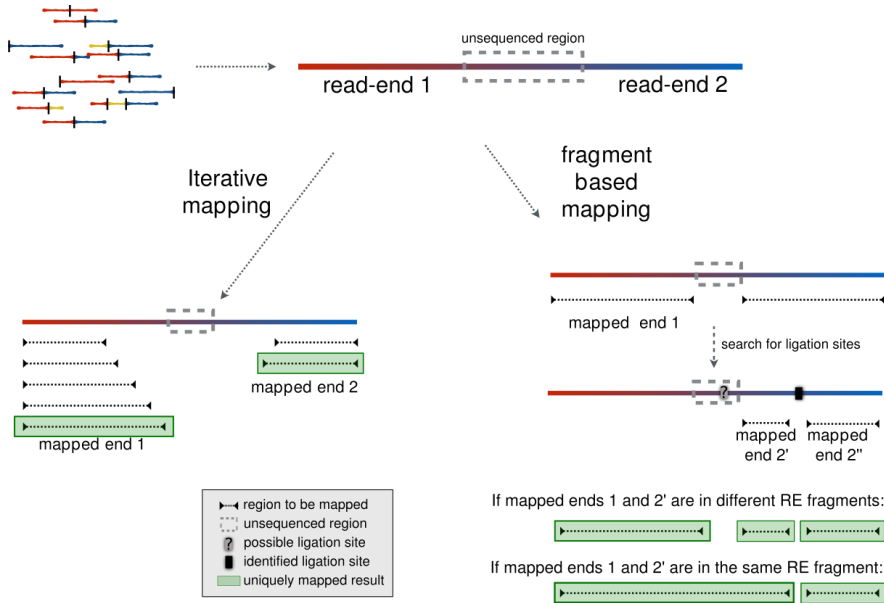


Figure 2.5. Mapping strategies implemented in TADbit. Taken from [34]

2.4.2.3 Fragment-level filtering

Some biases and errors from the experiment can be detected and corrected computationally. These include (see **Figure 2.6**):

- **Self-circle:** when the two ends of the same restriction fragment are ligated. It is identified when both read-ends map to the same fragment in opposed orientation.
- **Dangling-end:** when a fragment was not ligated. Identified in reads where the two sides map to the same restriction fragment in facing orientation.
- **Error:** when both sides of the read map to the same restriction fragment in the same orientation.
- **PCR artefacts or duplicated:** when the two reads have the same start position, mapped length, and strand, only one copy is kept.
- **Random breaks:** produced by non-canonical enzyme activity or random physical cleavage. They are detected when the distance from the read start

in any read-end and the restriction enzyme cut site is larger than a given threshold.

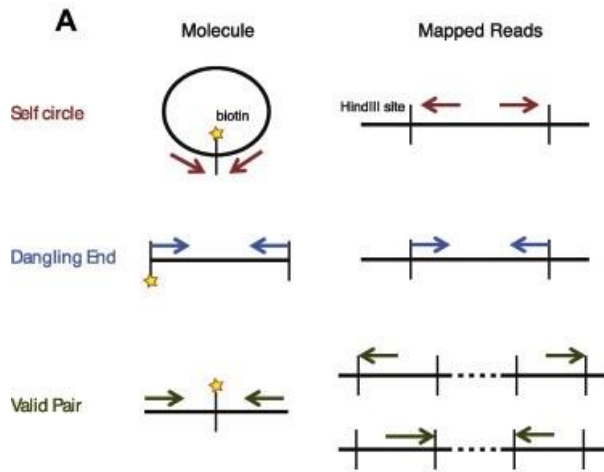


Figure 2.6. Fragment filtering in Hi-C data. Identification of molecule type in the mapped reads to discard artifacts based on their orientations relative to the restriction sites. Adapted from [35]

2.4.2.4 Bin-level filtering and normalization

The filtered fragments are binned at a user-specified resolution and summarized in a contact matrix where each cell represents the number of contacts identified between the two corresponding bins. Contact matrices are cleaned before normalization by removing columns with zero counts and those with less contacts than a given threshold. Normalization is based on the ICE (iterative correction and eigenvector decomposition) [36] and corrects for several sequence biases such as GC content or restriction site density. It iteratively balances the total counts of all bins, giving equal visibility to all genomic loci.

2.4.3 Identification of differential interactions

We used diffHiC [37], an R package to assess whether the interactions between pairs of loci significantly differ between two experimental conditions. It can model biological variability through quasi-likelihood (QL) methods considering the information of replicas. The counts y_{bi} of each bin pair b in the contact matrix of experimental sample i , are modeled using a Generalized Linear Model defined by

$$E(y_{bi}) = \mu_{bi} = \sum_{j=1}^p x_{ij}\beta_{bj} + o_{bi} \quad (2.5)$$

where x_{ij} are the elements of the design matrix that specify the experimental conditions of each sample and β_{bj} the corresponding unknown effects. The offsets o_{bi} represent normalization factors, for instance for the sequencing depth. The variability of each bin pair is given by

$$V(y_{bi}) = \sigma_b^2(\mu_{bi} + \phi_b \mu_{bi}^2) \quad (2.6)$$

where σ_b^2 is the QL dispersion parameter and ϕ_b the Negative Binomial dispersion. With this model, a QL F-test can be applied to each bin pair obtaining the fold change (FC) and the false discovery rate (FDR) correction for multiple testing of the obtained p-values. Then, significantly increasing interactions were defined as those bin pairs with $FDR < 0.5$ and $\log FC > 1$, and significantly decreasing interactions if the $FDR < 0.5$ and $\log FC < -1$.

2.4.4 Visualization

The statistically significant differential interactions can be visualized in a Circos plot [38], where the interactions are displayed as links joining the two bins that interact on a circular ideogram layout. Other annotations of genomic features can also be displayed for the genomic regions in the plot. We generated Circos plots for differentially increasing and decreasing interactions in *cis* and in *trans*.

Normalized contact matrices were transformed into Binary Upper Triangular MatRix (BUTLR) file format, using BUTLRtools (<https://github.com/yuelab/BUTLRTools>), suited for 3D Genome Browser (<http://3dgenome.org>) to visualize contact maps together with genome annotations [39]. Other genomic features can be included in the visualization to help in the interpretation of results (see **Figure 2.7**).

Another widely used software for visualizing Hi-C contact matrices is Juicebox [40]. It allows to display inter and intra chromosomal interaction matrices, interactively zooming and adjusting the binning level. Additionally, two experimental matrices

can be compared (displaying the ratio or the difference), 2D information such as TADs or loops can be displayed on top of the matrix and 1D data can also be added to the genomic axes.

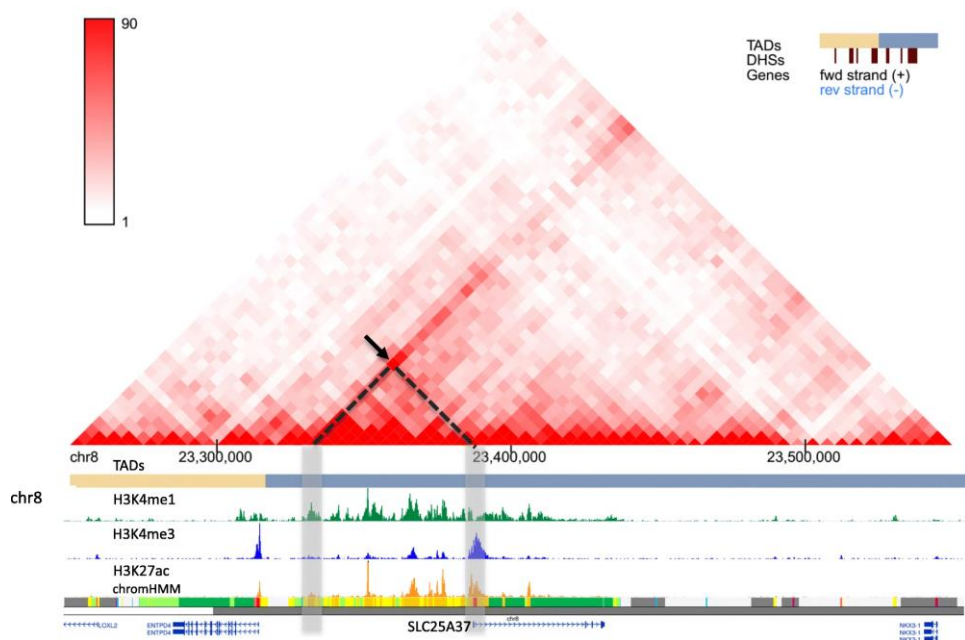


Figure 2.7. Visualization of a Hi-C contact matrix from K562 cells at 5kb resolution in the 3D Genome Browser. The intensity of the color (red) is based on the contact frequency between every bin pair. TADs are marked as yellow and blue bars. Histone marks (H3K4me1, H3K4me3 and H3K27ac), chromHMM chromatin types and gene positions are marked for every locus in the contact matrix. Adapted from [39]

2.5 DNA methylation

CpG methylation is important in gene regulation through different mechanisms. Methylation of cytosines in CpG steps changes the stiffness of DNA and therefore alters the nucleosome stability and the transcription factor binding affinity intrinsic to a given sequence [41]. This section presents some experimental techniques employed for the investigation of genome-wide DNA methylation.

2.5.1 Whole-genome bisulfite sequencing

Whole-genome bisulfite sequencing (WGBS) is a next-generation sequencing technique to detect the position of methylated cytosines (5mC) in a genome at single-

nucleotide resolution. It has allowed the analysis of cytosine methylation patterns in a wide range of organisms and cell types [42]. The DNA is treated with sodium bisulfate, causing unmethylated cytosines to be transformed to uraciles whereas methylated cytosines are not modified. Then, the treated samples are sequenced and the unmethylated cytosines are read as thymines from the polymerase chain reaction (PCR) amplification. Comparing the obtained reads with the untreated genome, the mismatches between C and T upon treatment correspond to the unmethylated cytosines and the matching C's correspond to methylated sites (see **Figure 2.8**).



Figure 2.8. Conversion of cytosines after treatment with bisulfite. Methylated cytosines (in red) are not modified upon treatment while unmethylated cytosines (in blue) are transformed to uraciles and then read as thymines after PCR amplification. Taken from [43]

In this work, the WGBS reads were processed using the gemBS pipeline v3.0 [42]. Reads with MAPQ scores < 20 or mapping to the same start and end points on the genome were filtered out. The first 5 bases from each read were trimmed before the variant and methylation calling step to avoid artifacts due to end repair. For each sample, CpG sites were selected where both bases were called with a Phred score of at least 20, corresponding to an estimated genotype error level of $\leq 1\%$. To exclude repetitive regions, loci with >500x coverage depth were excluded. From the successfully aligned reads, the methylation level of each CpG was computed as the ratio between the number of reads with an unconverted cytosine over the total number of reads (either with cytosine or thymine at that position).

2.5.2 Nanopore sequencing

As mentioned above, a disadvantage of WGBS is the inability to map sequenced reads in repetitive genomic regions. Additionally, it is not possible to determine whether two methylated cytosines separated by more than the fragment length occur in the same DNA molecule or come from different cells. Using Oxford Nanopore

Technology (ONT), much longer read lengths (>10kbp) can be obtained to identify methylated cytosines. Hence, it is possible to study the methylation at repetitive regions from the nanopore reads as well as the correlation of methylation at multiple CpG sites on the same DNA molecule.

2.6 ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is employed for the experimental study of protein-DNA interactions (transcription factors, histone modifications, RNA polymerase, etc.). It combines chromatin immunoprecipitation, where an antibody selects the regions bound by the target protein, and massively parallel sequencing for the detection of the binding sites.

The protocol starts with cross-linking of protein and DNA by treating cells with formaldehyde, to fix the position of the interaction. Then, the DNA is sheared with sonication or MNase to obtain short fragments that are then immunoprecipitated with the antibody specific to the protein of interest. The cross-linking in the selected DNA-protein complexes is then reversed and the DNA is purified and sequenced after size selection (typically fragment length ranges between 150 and 300 bp).

The experimental protocol presents some biases [44] such as the specificity of the antibody or the uneven fragmentation in open or closed chromatin. To account for the effect of these biases, it is important to include a control experiment. Three main types of control samples can be included:

- Input DNA (IP): before immunoprecipitation, a portion of the cross-linked and sheared fragments are selected.
- Mock IP DNA: the sample is immunoprecipitated without antibodies.
- DNA from nonspecific IP: the sample is immunoprecipitated using an antibody against a protein that does not bind to DNA and is not involved in chromatin modification, such as immunoglobulin G.

In this work, the sequenced reads were computationally analyzed using tools available in the Galaxy web platform [45]. First, reads were mapped to the reference

genome using BWA aligner [46]. Non-uniquely mapped reads were removed based on the mapping quality scores [47]. The coverage of mapped reads per base pair was then computed genome-wide and peak calling with MACS2 [48] performed to detect the protein binding regions, correcting the signal with the control samples.

2.7 RNA-seq

RNA sequencing (RNA-seq) is employed to quantify expression in a given transcriptome using next generation sequencing. The population of RNA in a sample is reverse transcribed into complementary DNA (cDNA) and adaptors are attached to both ends of the fragments. The library is high-throughput sequenced following manufacturer protocols. In the experiments analyzed in this thesis, TruSeq™ RNA Sample Prep Kit v2 (Illumina Inc.) was used to paired-end sequence the fragments with a read length of 2x76bp. Images analysis, base calling and quality scoring of the data was performed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and FASTQ sequence files were generated with CASAVA.

RNA-seq reads were aligned to the reference genome using GEM mapper [33] allowing for split maps. Genes were quantified using Flux-Capacitor [49], obtaining a table of the number of reads per gene in each sample. Comparison of expression levels between samples requires normalization of the gene counts, since library sizes can be different. The data was normalized by the trimmed mean of M-values (TMM) method of the edgeR software [50], which considers the possible differences in RNA distribution that might appear when changing experimental conditions (e.g. under stress conditions a specific set of genes might severely increase their expression levels). The normalized values were used to perform differential expression analysis using the "robust" version of the edgeR R package [51], which removes the bias from outliers while preserving high power detecting significant changes in expression.

Bibliography for Chapter 2

- [1] P. D. Dans *et al.*, "The physical properties of B-DNA beyond Calladine-Dickerson rules."
- [2] X.-J. Lu and W. K. Olson, "3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures," *Nucleic Acids Res.*, vol. 31, no. 17, pp. 5108–5121, Sep. 2003.
- [3] F. Lankaš, J. Šponer, J. Langowski, and T. E. Cheatham, "DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations," *Biophys. J.*, vol. 85, no. 5, pp. 2872–2883, Nov. 2003.
- [4] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes," *Proc. Natl. Acad. Sci.*, vol. 95, no. 19, pp. 11163–11168, Sep. 1998.
- [5] V. Matys, "TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D108–D110, Jan. 2006.
- [6] A. Mathelier *et al.*, "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles," *Nucleic Acids Res.*, p. gkv1176, Nov. 2015.
- [7] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat. Rev. Genet.*, vol. 5, no. 4, pp. 276–287, Apr. 2004.
- [8] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy, "Biostrings: String objects representing biological sequences, and matching algorithms." R package version 2.34.0., 2015.
- [9] B. Sollner-Webb and G. Felsenfeld, "Comparison of the digestion of nuclei and chromatin by staphylococcal nuclease," *Biochemistry*, vol. 14, no. 13, pp. 2915–2920, Jul. 1975.
- [10] R. Axel, "Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease," *Biochemistry*, vol. 14, no. 13, pp. 2921–2925, Jul. 1975.
- [11] W. K. M. Lai and B. F. Pugh, "Understanding nucleosome dynamics and their links to gene expression and DNA replication," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 9, pp. 548–562, May 2017.
- [12] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom, "A map of nucleosome positions in yeast at base-pair resolution," *Nature*, vol. 486, no. 7404, pp. 496–501, Jun. 2012.
- [13] R. V. Chereji, S. Ramachandran, T. D. Bryson, and S. Henikoff, "Precise genome-wide mapping of single nucleosomes and linkers in vivo," *Genome Biol.*, vol. 19, no. 1, Dec. 2018.
- [14] Ö. Deniz, O. Flores, F. Battistini, A. Pérez, M. Soler-López, and M. Orozco, "Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast," *BMC Genomics*, vol. 12, no. 1, Dec. 2011.
- [15] G. Gutiérrez *et al.*, "Subtracting the sequence bias from partially digested MNase-seq data reveals a general contribution of TFIIS to nucleosome positioning," *Epigenetics Chromatin*, vol. 10, no. 1, Dec. 2017.
- [16] O. Flores, O. Deniz, M. Soler-López, and M. Orozco, "Fuzziness and noise in nucleosomal architecture," *Nucleic Acids Res.*, vol. 42, no. 8, pp. 4934–4946, Apr. 2014.
- [17] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.
- [18] E. Planet, C. S.-O. Attolini, O. Reina, O. Flores, and D. Rossell, "htSeqTools: high-throughput sequencing quality control, processing and visualization in R," *Bioinformatics*, vol. 28, no. 4, pp. 589–590, Feb. 2012.

- [19] O. Flores and M. Orozco, “nucleR: a package for non-parametric nucleosome positioning,” *Bioinformatics*, vol. 27, no. 15, pp. 2149–2150, Aug. 2011.
- [20] D. Buitrago *et al.*, “Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning,” *Nucleic Acids Res.*, p. gkz759, Aug. 2019.
- [21] M. Simonis *et al.*, “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C),” *Nat. Genet.*, vol. 38, no. 11, pp. 1348–1354, Nov. 2006.
- [22] Z. Zhao *et al.*, “Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions,” *Nat. Genet.*, vol. 38, no. 11, pp. 1341–1347, Nov. 2006.
- [23] J. Dostie *et al.*, “Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements,” *Genome Res.*, vol. 16, no. 10, pp. 1299–1309, Oct. 2006.
- [24] E. Lieberman-Aiden *et al.*, “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome,” *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009.
- [25] J. R. Hughes *et al.*, “Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment,” *Nat. Genet.*, vol. 46, no. 2, pp. 205–212, Feb. 2014.
- [26] S. Schoenfelder *et al.*, “The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements,” *Genome Res.*, vol. 25, no. 4, pp. 582–597, Apr. 2015.
- [27] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, “Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C,” *Cell*, vol. 162, no. 1, pp. 108–119, Jul. 2015.
- [28] M. Forcato, C. Nicoletti, K. Pal, C. M. Livi, F. Ferrari, and S. Bicciato, “Comparison of computational methods for Hi-C data analysis,” *Nat. Methods*, vol. 14, no. 7, pp. 679–685, Jun. 2017.
- [29] P. Hansen *et al.*, “Computational Processing and Quality Control of Hi-C, Capture Hi-C and Capture-C Data,” *Genes*, vol. 10, no. 7, p. 548, Jul. 2019.
- [30] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. J. Filion, and M. A. Marti-Renom, “Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors,” *PLOS Comput. Biol.*, vol. 13, no. 7, p. e1005665, Jul. 2017.
- [31] S. Andrews, “FastQC: A quality control tool for high throughput sequence data,” 2010. [Online]. Available: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [32] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, “Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment,” *Genome Res.*, vol. 8, no. 3, pp. 175–185, Mar. 1998.
- [33] S. Marco-Sola, M. Sammeth, R. Guigó, and P. Ribeca, “The GEM mapper: fast, accurate and versatile alignment by filtration,” *Nat. Methods*, vol. 9, no. 12, pp. 1185–1188, Oct. 2012.
- [34] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. J. Filion, and M. A. Marti-Renom, “Iterative vs fragment-based mapping,” *TADbit Tutorial*, 2017. [Online]. Available: https://3dgenomes.github.io/TADbit/tutorial/tutorial_4-Mapping.html. [Accessed: 09-Sep-2019].
- [35] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, “Hi-C: A comprehensive technique to capture the conformation of genomes,” *Methods*, vol. 58, no. 3, pp. 268–276, Nov. 2012.
- [36] M. Imakaev *et al.*, “Iterative correction of Hi-C data reveals hallmarks of chromosome organization,” *Nat. Methods*, vol. 9, no. 10, pp. 999–1003, Sep. 2012.
- [37] A. T. L. Lun and G. K. Smyth, “diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data,” *BMC Bioinformatics*, vol. 16, no. 1, Dec. 2015.

- [38] M. Krzywinski *et al.*, "Circos: An information aesthetic for comparative genomics," *Genome Res.*, vol. 19, no. 9, pp. 1639–1645, Sep. 2009.
- [39] Y. Wang *et al.*, "The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions," *Genome Biol.*, vol. 19, no. 1, p. 151, Dec. 2018.
- [40] N. C. Durand *et al.*, "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments," *Cell Syst.*, vol. 3, no. 1, pp. 95–98, Jul. 2016.
- [41] G. Portella, F. Battistini, and M. Orozco, "Understanding the Connection between Epigenetic DNA Methylation and Nucleosome Positioning from Computer Simulations," *PLoS Comput. Biol.*, vol. 9, no. 11, p. e1003354, Nov. 2013.
- [42] A. Merkel *et al.*, "GEMBS – high through-put processing for DNA methylation data from Whole Genome Bisulfite Sequencing (WGBS)," *Bioinformatics*, preprint, Oct. 2017.
- [43] C. Grehl, M. Kuhlmann, C. Becker, B. Glaser, and I. Grosse, "How to Design a Whole-Genome Bisulfite Sequencing Experiment," *Epigenomes*, vol. 2, no. 4, p. 21, Dec. 2018.
- [44] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 669–680, Oct. 2009.
- [45] E. Afgan *et al.*, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W537–W544, Jul. 2018.
- [46] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinforma. Oxf. Engl.*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009.
- [47] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, Nov. 2008.
- [48] Y. Zhang *et al.*, "Model-based Analysis of ChIP-Seq (MACS)," *Genome Biol.*, vol. 9, no. 9, p. R137, 2008.
- [49] S. B. Montgomery *et al.*, "Transcriptome genetics using second generation sequencing in a Caucasian population," *Nature*, vol. 464, no. 7289, pp. 773–777, Apr. 2010.
- [50] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biol.*, vol. 11, no. 3, p. R25, 2010.
- [51] X. Zhou, H. Lindsay, and M. D. Robinson, "Robustly detecting differential expression in RNA sequencing data using observation weights," *Nucleic Acids Res.*, vol. 42, no. 11, pp. e91–e91, Jun. 2014.

Chapter 3 . Sequence dependent DNA flexibility and protein recognition

The shape of the DNA duplex was first described from diffraction data several years ago [1], [2], and since then many experimental techniques have completed our view of how DNA duplex is under physiological conditions: a very flexible and polymorphic duplex [3], [4] which can adopt different conformations depending on the sequence, environment and presence of DNA-binding proteins [5]–[8]. Such an intrinsic polymorphism is crucial for its functionality.

As explained in Chapter 2, the base pair step geometry can be represented by a set of six helical parameters describing translations and rotations of one given base pair with respect to the neighboring one. The DNA flexibility was evident from the structural variability observed for the same complex in different crystals [9], which suggested that flexibility could be simulated by using an harmonic model with stiffness constants derived from the observed variability in the distributions. This work was posteriorly extended retrieving the helical coordinates from trajectories obtained from molecular dynamics (MD) simulations [10], [11], which helped to solve the problem of the lack of experimental data. Nowadays, the developments in atomistic MD simulations and accurate forcefields [12] allowed obtaining long reliable trajectories and sampling the conformational space of different DNA sequences, and revealed that the dinucleotide-model is not sufficient for describing the high flexibility of DNA molecules [13], [14], and that at least a tetranucleotide model should be used [13], [14].

Extending the analysis to the nearest neighbors of each dinucleotide, a tetramer model of sequence dependence has been studied on a large collection of trajectories from MD simulations from the Ascona B-DNA Consortium (ABC, <https://bisi.ibcp.fr/ABC>) and the BigNASim database [15]. These studies confirmed that, for most tetramers, considering only the nearest neighbors is sufficient for describing their structure and flexibility, but a few tetranucleotides exhibit highly polymorphic behavior and dependence on the sequence context beyond the tetramer level. One of these tetramers, CTAG, has been extensively studied in the work *Modulation of the helical properties of DNA: next-to-nearest neighbor effects and beyond*, that is attached below as part of this thesis, where we found evidence of the unique structural properties of this sequence, which might confer special flexibility related to its particular location in the genome and its low mutation rate.

The tetramer base (pseudo) harmonic model (or its extension to the hexamer level; see below) can be used to describe the energetic cost of deforming a DNA structure. Particularly, the model can be used to determine the ease in which a given DNA sequence can be deformed to adopt the conformation when bound to an effector protein, and accordingly can provide information on the sequence-preference of a given DNA interacting protein. In this thesis, the structural differences between the free and the protein-bound DNA were studied and the energetic cost related to the structural changes of the DNA to adopt the conformation in the protein complex were calculated (see publication *How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition*). Using statistical tests to analyze the helical motions, it was found that a large percentage of the DNA sequences studied can spontaneously sample the bioactive conformation, while a small percentage is highly distorted by the protein binding, due to strong non-harmonic deformations such as base opening.

3.1 Modulation of the helical properties of DNA: next-to-nearest neighbor effects and beyond

Our group studied the physical properties of DNA sequences corresponding to the ten possible base pair steps in all the possible tetramer environments [16]. The study showed that while several base pair step helical parameters can sample different configurations along the MD simulations, having a population of values that correspond to a normal distribution, some deviate from normality and have multimodal distributions. Therefore, more general models including information beyond the dinucleotide level are required. Moreover, it was found that some tetramers were ultra-flexible, and their conformation might be modulated by effects beyond the tetramer level which are rare for the rest. In this work, structural analysis of one of these highly flexible tetramers, CTAG, is presented as well as a genomic analysis of its prevalence in different species.

We analyzed 40 different sequence contexts containing CTAG in a central position carefully selected to cover all the possible hexamers. First, examining the distributions of helical parameters of the central TA base pair step retrieved from individual trajectories, we observed deviations from the normal distributions showing multimodal densities for some parameters (shift, slide and twist). Important differences in their distributions were also detected when different sequence contexts were considered. Our analysis suggests that the multimodality can be explained by sequence effects beyond the nearest neighbors, at the hexamer or even octamer level.

Additionally, data mining of experimental structural data, obtained from the Protein Data Bank (PDB, [17]), was performed in order to validate our conclusions. Although the obtained data is scarce, limiting the generality of the conclusions, the results are in line with our MD-based observations, showing that the multimodality in the distributions is not an artefact of the force-field used in the MD simulations.

Furthermore, we investigated whether the high flexibility of CTAG tetramer might confer some specific functionality in different eukaryotic genomes. Interestingly, this tetramer sequence is unfrequently found in several genomes analyzed. We

evaluated whether its low frequency was because it contains one of the stop codons (TAG) and we concluded that it is not the case, since the frequency of other tetramers that contain this stop codon is on the average compared to all possible tetramers. We found out that this very peculiar tetramer is underpopulated along the genome and preferentially found in intergenic regions, and unfrequently detected in coding regions. Moreover, investigating data collected for different cancer types, we observed that its mutation frequency is low compared to other tetramers.

Publication:

Alexandra Balaceanu, Diana Buitrago, Jurgen Walther, Adam Hospital, Pablo D. Dans. and Modesto Orozco. (2019). Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond. *Nucleic Acids Research*, 47, 4418–4430. <https://doi.org/10.1093/nar/gkz255>.

Supplementary material for this article can be found in the **Annex I**.

4418–4430 *Nucleic Acids Research*, 2019, Vol. 47, No. 9
doi: 10.1093/nar/gkz255

Published online 8 April 2019

Modulation of the helical properties of DNA: next-to-nearest neighbour effects and beyond

Alexandra Balaceanu¹, Diana Buitrago¹, Jürgen Walther¹, Adam Hospital¹, Pablo D. Dans¹ and Modesto Orozco^{1,2,*}

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain and ²Department of Biochemistry and Biomedicine, University of Barcelona, 08028 Barcelona, Spain

Received February 04, 2019; Revised March 23, 2019; Editorial Decision March 28, 2019; Accepted March 30, 2019

ABSTRACT

We used extensive molecular dynamics simulations to study the structural and dynamic properties of the central d(TpA) step in the highly polymorphic d(CpTpApG) tetranucleotide. Contrary to the assumption of the dinucleotide-model and its nearest neighbours (tetranucleotide-model), the properties of the central d(TpA) step change quite significantly dependent on the next-to-nearest (hexanucleotide) sequence context and in a few cases are modulated by even remote neighbours (beyond next-to-nearest from the central TpA). Our results highlight the existence of previously undescribed dynamical mechanisms for the transmission of structural information into the DNA and demonstrate the existence of certain sequences with special physical properties that can impact on the global DNA structure and dynamics.

INTRODUCTION

Early structural models of DNA derived from fibre diffraction data provide a static and averaged picture of the double helix (1–3), which despite its simplicity was sufficient to represent the general shape of DNA in physiological conditions. However, as more accurate structural techniques appeared, the intrinsic polymorphism of double-stranded DNA become evident (4–7) as significantly different conformations were described depending on the sequence, the environment or the presence of ligands (8–11). Six decades after the development of the first duplex models, we understand that DNA as a flexible and polymorphic molecule is able to sample a wide range of helical geometries, thanks to a complex choreography of backbone rearrangements, which allows the conformational changes required for DNA functionality (11–19).

Attempts to determine the principles relating sequence and structure originated in the eighties when by process-

ing the scarce experimental data available, Calladine *et al.* (20), developed a series of heuristic rules relating sequence with some structural characteristics of DNA (21,22). In the late nineties (23), Olson *et al.* developed a complete set of parameters defining the expected distribution of helical parameters of the 10 unique base pair steps (bps). Parameters were derived from the analysis of the available crystal data on DNA–protein complexes and provided information not only on the equilibrium geometry but also on the expected flexibility of the bps (extracted from the variability of the same bps in different crystals). Twenty years after their generation, Olson–Zhurkin parameters are still used to represent DNA by means of helical mesoscopic descriptors. However, we cannot ignore the strong assumptions involved in their derivation: (i) the ensemble of configurations obtained from the analysis of crystal structures should define a densely populated Gaussian distribution; (ii) a dinucleotide (step) model is enough to represent DNA sequence variability, i.e. the helical geometry can be decomposed at the bps level; (iii) conformational variability found in structures in PDB should exclusively depend on the flexibility of the step and finally (iv) binding of a protein should not introduce anharmonic distortions in the duplex geometry.

The eruption of atomistic molecular dynamics (MD) simulations gave the community an alternative source of parameters to describe DNA structure and flexibility. Compared with results derived from the analysis of experimental structures, the MD-based ones are more robust as they are obtained from processing an extremely large number of snapshots, and provide information on flexibility that is not contaminated by the presence of ligands, crystal lattice or any other environmental artifacts. As a major caveat, MD-derived descriptions of DNA properties are dependent on the length of trajectories as well as on the quality of the force field parameters used to describe DNA interactions. Thus, early attempts to describe DNA from multi-nanosecond trajectories led to artefactual results due to a previously unknown error of the most used force field at that time (24). A newer force field (25) and higher computa-

*To whom correspondence should be addressed. Tel: +34 93 40 37156; Email: modesto.orozco@irbbarcelona.org

tional capabilities provided descriptions of DNA properties that were more reasonable, but still far from the required accuracy (12,26,27). The availability of the highly accurate PARMBSC1 force field (28,29) and the development of new MD codes taking advantage of a new generation of computers (30–33) provide the community with the possibility to derive reliable representation of the sequence-dependent physical properties of DNA from the analysis of microsecond long trajectories collected under highly controlled simulation conditions.

Results collected by the Ascona B-DNA Consortium (34–37) revealed two major findings that challenged current models of DNA flexibility. First, the dinucleotide-model is insufficient to describe DNA flexibility, as the variability in bps parameters depending on tetranucleotide environment can be more pronounced than the variability found when comparing different bps for a given tetranucleotide context. Second, several distributions of helical parameters considering the nearest neighbours deviate from normality and a part of them are in fact multi-modal, which means that the physical properties of such tetranucleotides cannot be represented by a single set of elastic parameters (equilibrium values and associated stiffness). Analysis of MD data revealed that the changes between substates happen towards a series of coordinated changes along the backbone (17,37,38), where unusual H-bond interactions and subtle changes in the solvent environment play a key role (18,39). The analysis of ABC data and of additional trajectories stored in our BigNASim database (40) suggested that a nearest neighbour-based model was in general sufficient to derive transferable descriptors of DNA structure and flexibility, but a few exceptions to this general rule emerged; the clearest one is the d(CpTpApG) tetranucleotide (in the following CTAG): a very polymorphic stretch of DNA, with 50% G-C content, for which results were significantly different depending on the simulation. The structural peculiarities of TpA steps have been qualitatively pointed out in the past by analysing a small number of experimental structures, especially when immersed in short A-tracks (41,42).

We present here a detailed analysis of CTAG in different sequence contexts. Results demonstrate that next-to-nearest effects modulate the geometrical properties of the central d(TpA) step. Such structural effects are very visible when hexanucleotides are considered, but quite surprisingly extend beyond the next-to-nearest level, indicating the existence of a complex mechanism of information transfer across DNA through the coordinated backbone and base movements.

MATERIALS AND METHODS

The choice of sequences and the simulation details

The systematic study of sequence-dependent effects beyond the tetranucleotide level has been to date impossible, due to the huge number of sequences that need to be considered. For example, the study of all hexanucleotides would require the simulation of 2,080 sequences, while to consider all octanucleotides 32,826 sequence combinations are needed. Fortunately, the analysis of ABC simulations where tetranucleotides appear in different molecular environments suggests that sequences effects beyond the

tetranucleotide are rare, and if they exist, are localized in certain ultra-flexible sequences. We focused our interest here in one of the most flexible tetranucleotide: CTAG. Thus, we built a library of 40 different sequences covering the entire hexanucleotide space (XpCpTpApGpX) as well as all possible pyrimidine(Y)/purine(R) combinations at the octanucleotide level in several repeats (see Supplementary Methods). All the sequences were prepared using the leap module of AMBERTOOLS 16 (43) and standard ABC protocol (37). Accordingly, systems were built from Arnott's B-DNA average parameters, neutralizing the DNA with K⁺ ions, adding water (at least 10 Å of water separate DNA from the faces of the box) and extra 150 mM KCl. Systems were then optimized, thermalized and equilibrated before production (34,35). Water was represented with the SCP/E model (44), Smith-Dang parameters were used for ions (45–47) and the recent PARMBSC1 force field was considered to represent nucleic acids interactions (28). Trajectories (collected in the NPT ensemble $T = 298$ K, $P = 1$ atm) were extended from 0.5 μ s to up to 9 μ s. All simulations were performed with the pmemd.cuda code using periodic boundary conditions and Particle Mesh Ewald (31,48). Movements of hydrogen atoms were annihilated using SHAKE (49), which allowed us the use of a 2 fs integration step. All trajectories collected here are accessible through the MuG BigNASim portal (40): <https://mmb.irbbarcelona.org/BIGNASim/>

Analysis

Standard analysis was done using *cptraj* module of the AMBERTOOLS 16 package (43), the NAFlex server (50) CURVES+ and CANAL programs (51), following the standard ABC-conventions (37). The CANON module from Curves+ (38) was used to determine distributions of ion populations in curvilinear cylindrical coordinates for each snapshot of the simulations with respect to the instantaneous helical axis. Duplexes were named following the Watson strand (e.g. ATGG stands for (ATGG)-(CCAT)). The letters R, Y and X stand for a purine, a pyrimidine or any base respectively, while X:X and XX represent a base pair and base pair step, respectively. Base pairs flanking the CTAG were denoted using two dots to represent the central tetrad (e.g. R··Y). The normality and modality of the helical distributions were evaluated using Bayesian Information Criteria (52,53) and Helguero's theorem (54) as described elsewhere (12). Classification of the torsional states of the different rotatable bonds in the DNA backbone was done using standard criteria (55). Correlations between different torsions were determined by circular correlation analysis (see Supplementary Methods for additional details). The meta-trajectory analysis was used to define the global characteristic of the d(TpA) essential deformation space. With this purpose, the 40 individual trajectories were grouped and subjected to principal component analysis (56,57) in the helical space of the central d(TpA) step after Lankaš' normalization of the different rotational and translational degrees of freedom (58). The essential dynamics of the central d(TpA) step is then used to define the set of key movements explaining the global deformation at the d(TpA) step. The distributions of the four informative bps deformations were subjected to detailed analysis (see Supplementary Method

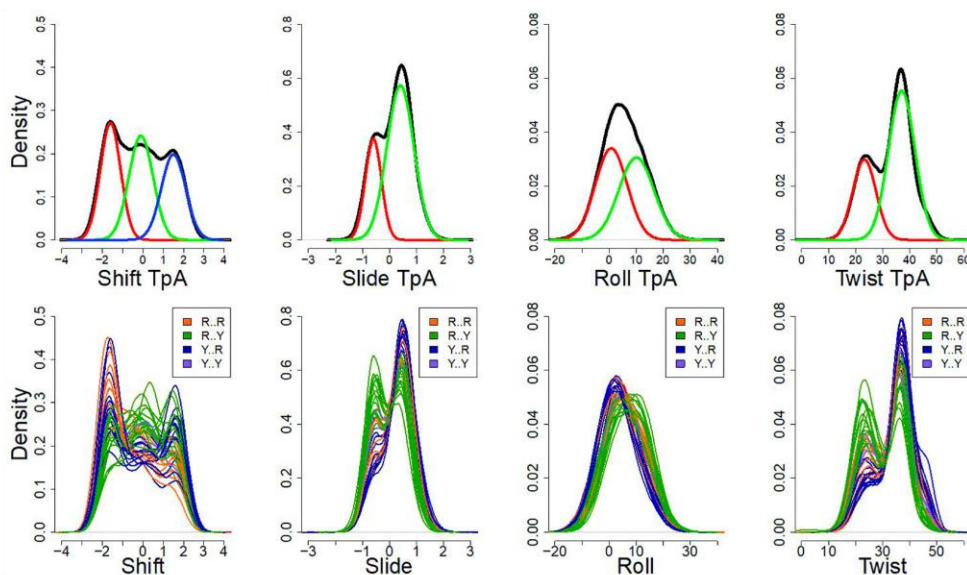
4420 *Nucleic Acids Research*, 2019, Vol. 47, No. 9

Figure 1. Normalized frequencies of those bps helical parameters found to be bi-normal and tri-normal according to the BIC analysis. First row: Density obtained from the meta-trajectory (black line), and the BIC decomposition in two Gaussians (slide, roll and twist: red and green lines) or in three Gaussians (shift: red, green and blue lines). Second row: Overlapped density of the shift, slide, roll and twist parameters at the central TpA step of the 40 sequences studied (see Supplementary Table S1).

for additional details). Comparison and clustering of the individual trajectories of the central d(TpA) for the 40 sequences studied (all with a common CTAG central tetranucleotide) were done using symmetrized Kullback-Leibler (KL) divergences (58) followed by hierarchical cluster analysis using Ward's clustering criterion (59), where the dissimilarities are squared before cluster updating (60), using as descriptive variable the six distinguished helical variables detected by the PCA of the meta-trajectory (see Supplementary Methods for additional details). The clusters obtained in this manner were subsequently analysed in detail, further highlighting the differences between their individual accessible helical spaces. Ion analysis was performed as described elsewhere (18,38) to unravel the connections between the binding of cations on the DNA and its mechanistic properties. Stacking strengths were followed by geometrical criteria for the central dinucleotide in the meta-trajectory filtered by the three main states in helical space, as described in detail in Supplementary Methods. Structural database analysis was done using all DNA structures containing the CTAG tetranucleotide. Genomic analysis was done to determine the prevalence of the CTAG tetranucleotide in different wild-type genomes and its resilience to mutation. Genomes of *Homo sapiens* (hg19), *Escherichia coli* (NC.000913.3) and *Saccharomyces cerevisiae* (sacCer3) were analysed. Occurrences of this tetranucleotide were then mapped, using Homer software (61), to the annotated regions of each organism obtained from UCSC and compared to the overall frequency of each annotation type. To

compute the resilience to mutation, the frequency of mutations for each tetranucleotide along the genome in 30 different cancer types (data from (62)) was determined normalizing by tetranucleotide occurrence along the genome. Single-nucleotide polymorphisms (SNPs) in the human genome were retrieved from Ensembl Variation database (63), and the number of SNPs per tetranucleotide was computed, normalizing by genome-wide tetranucleotide frequency.

RESULTS AND DISCUSSION

The CTAG shows dramatic and complex structural polymorphism

We collected trajectories for 40 oligonucleotides containing the CTAG tetranucleotide in a central position (see 'Materials and Methods' and Supplementary Table S1). All the trajectories were stable along time in the sub-microsecond timescale, sampling structures that fit well in the B-like double helical conformation. As suggested by the analysis of ABC-simulations (37), and of trajectories deposited in BigNASim, (40) CTAG is highly polymorphic as seen from clear bimodal distributions of some helical parameters. To check that the multi-peaked distributions were not artefacts due to limited sampling, we extended trajectories for selected tetranucleotides up to 9 μ s (Supplementary Table S1), tracing the changes in the distribution of helical parameters. The good convergence shown in Supplementary Figure S1 supports the robustness of our results and sug-

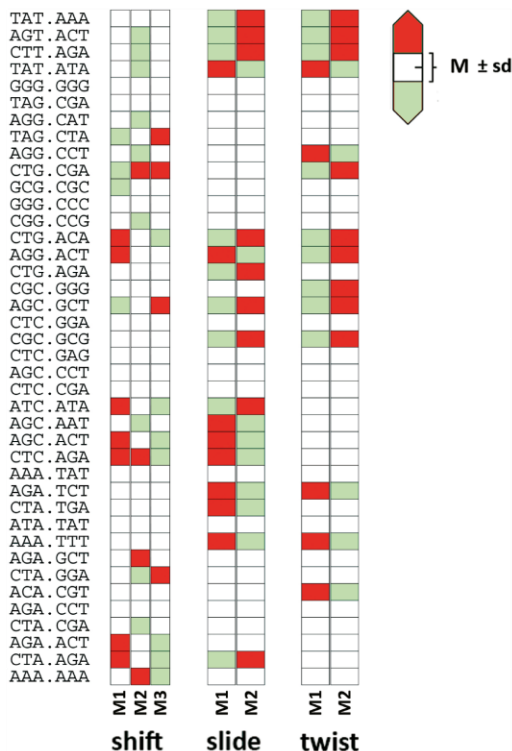


Figure 2. Relative propensities of the multi-modal bps helical coordinates of the central TpA in all 40 sequence contexts. Comparison to the global average propensities over all sequence contexts per component of the multi-modal distributions with standard deviations that reflect the variation of the propensity of each component amongst sequences. The propensity values were computed BIC analysis (see 'Materials and Methods' section and Supplementary Methods).

gests a fast dynamic of interchange of the different states (see 'Discussion' section below).

In order to obtain a global average picture of the conformational space accessible to the CTAG tetranucleotide, we joined the 40 individual trajectories (equal number of snapshots in all cases) to generate a meta-trajectory, which was then subjected to PCA and BIC analysis. Four base-parameters (the symmetric buckle and propeller twist of d(T·A) and d(A·T)) and four bps parameters at the central d(TpA) step (roll, twist, shift and slide) emerged as determinant to explain 60% of the variance in the meta-trajectory; Six of which were used for further analysis. As seen in the BIC analysis summarized in Figure 1, deviations from Gaussianity in the form of multi-peaked distributions are the main responsible for the structural polymorphisms detected at the bps level. Such deviations could in principle emerge from two different sources: (i) intrinsic multi-modality in the individual trajectories and (ii) indi-

vidual distributions (coming from the 40 sequences studied) are Gaussian, but they are centred at different average values. To analyse which is the real origin of the deviation from normality in meta-trajectories, we repeated the analysis for individual trajectories (Figure 1). Roll distributions were unimodal in all cases, but the position of the peak was displaced towards slightly higher values when the central tetranucleotide is surrounded by R at 5' and Y at 3' (i.e. RpCpTpApGpY hexanucleotides), leading to a bi-normal distribution of the meta-trajectory (see Figure 2). The situation is completely different for twist, slide and shift where bi- or even tri-modality (three peaks in the distribution) is clear for individual sequences (see Figure 2 and Supplementary Figure S2), with the different substates being sampled in a fast equilibrium along the time scale of the simulations (see examples in Supplementary Figure S3).

As shift distribution is tri-modal and twist and slide distributions are bi-modal, we could in principle expect 12 states. However, many of the combinations of twist, slide and shift substates are not possible, and in practice, only four states appear when meta-trajectory is projected in the twist-slide-shift 3D space (Figure 3). In fact, one of them (high twist/positive slide/zero shift; HPZ) is populated only in some of the simulations and has globally a reduced impact in the meta-trajectory ensemble, which is dominated by three main states (Figure 4): high twist/positive slide/negative shift (HPN); high twist/positive slide/positive shift (HPP) and low twist/negative slide/zero shift (LNZ). Experimental validation of the suggested polymorphisms is difficult as experimental structures are always averaged (i.e. assuming a normal unimodal distribution). However, plotting the scarce experimental data available for the CTAG tetranucleotide on the 2D population plots (shift-twist, shift-slide and twist-slide) derived from meta-trajectories provides an indirect, but strong support to our results. For example, the shift distribution is very narrow and centred around zero for low slide values, while when slide increases, larger values (either positive or negative) of shift are sampled, in perfect agreement with MD meta-trajectories. Similarly, low twist appears experimentally only in zero shift conformations, while high shift (either negative or positive) is found only in experimental structures with a high twist. Although the major discrepancies between MD and experiments seem to occur for the twist-shift plane, filtering the shift values according to low/high twist reconcile partially the matching between experiments and theory (Supplementary Figure S4). Finally, the twist-slide plot shows only two regions of high probability consistent with the same slide/twist correlation found experimentally (see Figure 3 and 'Discussion' section below).

Next-to-nearest dependence in central d(TpA) conformation

All the sequences studied here correspond to the same tetranucleotide, so a similar distribution of helical parameters at the central d(TpA) step could be expected. However, this is not the case as shown in selected examples in Supplementary Figure S2, where large differences in the distributions of helical coordinates for the d(TpA) step appear. Analysis of the trajectories (Figure 1) reveals that the origin

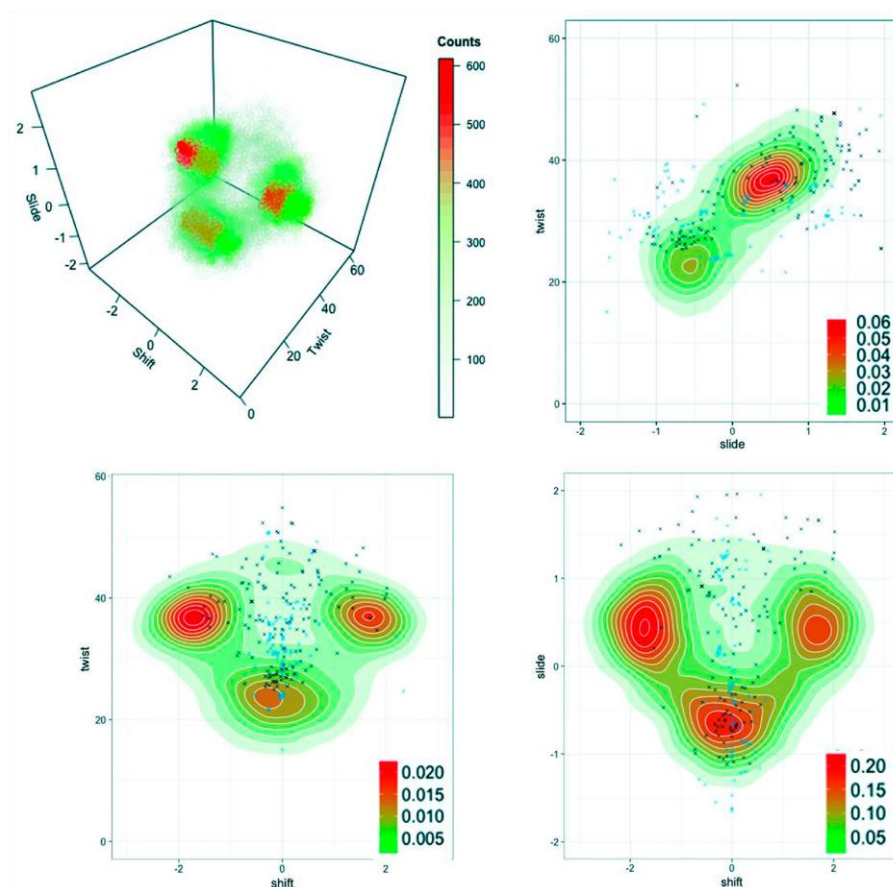
4422 *Nucleic Acids Research*, 2019, Vol. 47, No. 9

Figure 3. 3D and 2D counts in the shift, slide and twist planes from MD simulations at the central TpA step. In the 2D density plots, experimental structures from the PDB (see Supplementary Methods) were added as black crosses (protein–DNA complexes) or blue crosses (isolated DNA).

of the difference emerges from the different weights of the individual substates defining the global distributions (see a global summary in Figure 2). Moreover, we observe that the varying populations of these substates are a direct consequence of sequence context. To go deeper in the analysis of this hexanucleotide variability, we perform Kullback-Leibler (KL) analysis of the 40 trajectories in the 6D space defined from the PCA analysis as informative of the entire flexibility space of the helix (see above). Clustering analysis can be performed from the KL results to determine the similarity between sequences based on the dynamics of the central d(TpA) step and organized in the relational dendrogram (Figure 5), which clearly shows the presence of at least two major clusters. The first one is populated mainly by sequences where the central tetranucleotide is flanked by Y at 5' and R at 3', but also contains two 5'Y-3'Y sequences. The other cluster, the largest one, is subdivided

into three different subclusters, two of which are formed almost exclusively of sequences where the central tetranucleotide is surrounded by R at 5' and Y at 3'; finally, the last cluster corresponds to situations where the CTAG tetrad is surrounded by 5'R-3'R. Examples of prototypical distributions obtained for representative sequences in each cluster are shown in Supplementary Figure S5, which demonstrate that the hexanucleotide content has a non-negligible role in defining the properties of the central d(TpA) step in the CTAG tetranucleotide, a clear exception of the nearest neighbour model. Furthermore, the presence of some hexanucleotides in different clusters suggests that some couplings might be possible even beyond the next-to-nearest neighbour level (see below). The rules that govern the sampling of a given substate of the sequences in each cluster can be understood by analysing sequence-dependent stabilizing

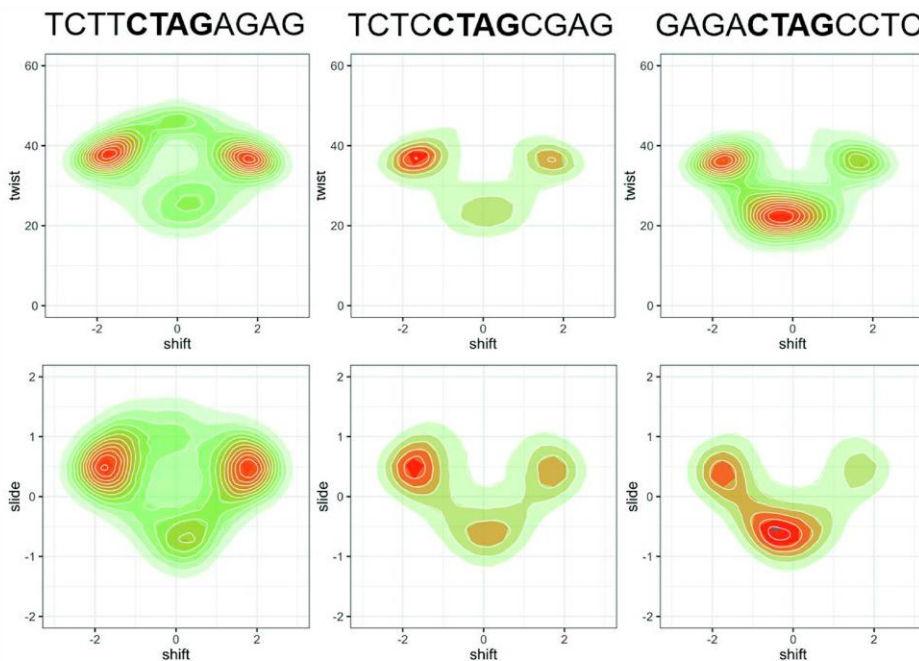


Figure 4. 2D density plots in the shift/twist and shift/slide planes at the central TpA step for three selected sequences.

factors that give rise to the characteristic distributions of helical parameters depicted in Supplementary Figure S5.

The existence of such effects implies that the motion of the central TpA step must be somehow connected to the distant base pairs. Mechanical information should travel from one site to the other to allow the TpA step to 'feel' its environment and respond in a different way according to the nature of the base pairs located almost half helical turn away. We were able to find a possible explanation based on the concerted and correlated movements of the backbone and bases, by first noting that the twist polymorphism at TpA was behaving as the better well-known YpR step: d(CpG) (18,34,37,39). The two possible twist substates (HT/LT) at the TpA step were connected to the backbone BI/BII polymorphism at the next GA junction (note that BI/BII interconversion is mainly governed by the ζ torsion). Furthermore, the BI/BII polymorphism at GpA is possible due to the formation of the intra C8H8-O3' h-bond and the shift polymorphism in the same junction (Figure 6A and B) (39). Similar results were found if looking to the correlation of twist at the central TpA step with the bps at the 5'-side (CpT). It is then clear that the main backbone polymorphism (BI/BII) is linked to the base polymorphisms, mainly to shift and twist (Supplementary Table S2) up to the next-to-nearest neighbours. The information travels through successive backbone and base polymorphisms, which are limited to some specific substates

due to DNA's cranksaft motion (Supplementary Table S2). This dynamically concerted movement of either (alone or in combination) shift/slide/twist step parameters and the ζ torsion could be appreciated from the Pearson correlation coefficients that clearly show a correlation/anti-correlation pattern in successive bps. Since intra-molecular CH-O h-bonds are mainly responsible for the information transfer between the backbone and the base (39) (with perhaps a small contribution from the known sugar pucker flexibility, see Supplementary Table S2), both backbone and base polymorphisms can be followed by looking only to the formation of those C8H8-O3' h-bonds in RpR and YpR steps, or C6H6-O3' h-bonds in RpY and YpY steps. The correlated/anti-correlated formation of these h-bonds away from the central TpA step clearly explains the transfer of mechanical information up to the next-to-nearest neighbours, and also beyond depending on the sequence (see 'Discussion' section below and Figure 6C). As a general rule, at the tetranucleotide level, the BII backbone state is significantly favoured at the 3' side on either strand (i.e. at GpA step). The correlations of backbone substates with the helical parameters at TpA paint a picture where negative shift is related to having more BI at the GpA of the Watson strand and more BII at GpA on the Crick strand, with positive shift being favoured in the exactly opposite situation. Additionally, the TpA can be found in a low twist state only when both 3' GpA junctions are in BII, while the simultane-

4424 Nucleic Acids Research, 2019, Vol. 47, No. 9

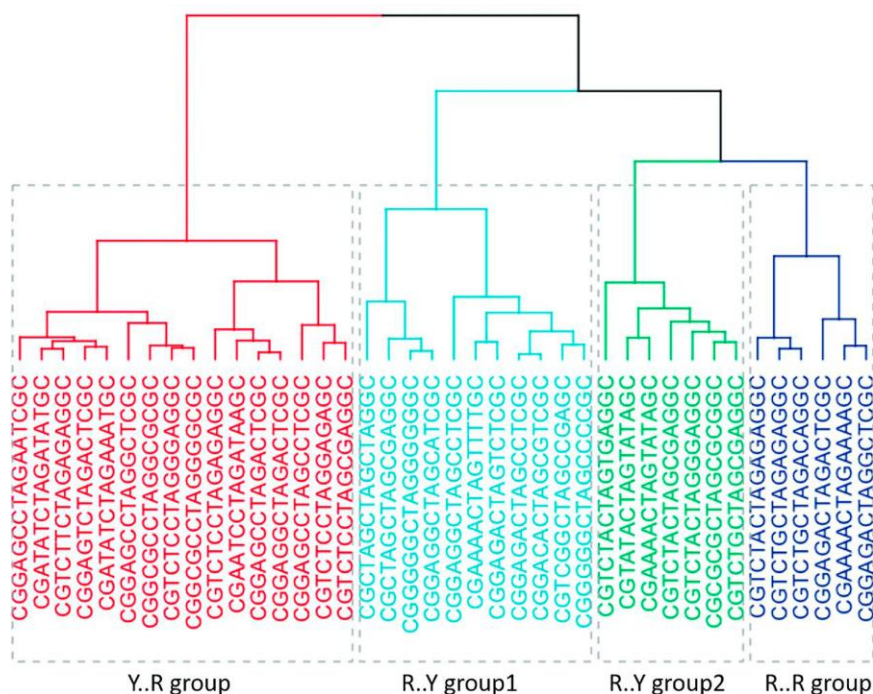


Figure 5. Dendrogram obtained from a hierarchical clustering method using Ward's criterion to classify the sequences. The distances were obtained from the symmetric Kullback-Leibler (KL) divergence in the space of six helical parameters: shift, slide and twist of TpA step, buckle and propeller of dT, and the buckle of dA (see Supplementary Methods).

ous BI state on both strands at GpA will promote high twist at TpA. The next-to-nearest context and sometimes more remote sequence effects can modulate the relative populations of BI/BII on the two strands, which in turn will affect the helical parameters at the central TpA. It's worth noting that the correlations between helical parameters in consecutive steps are mostly anti-correlations, and in general the global twist distribution of a tetra- or hexanucleotide segment can be perfectly described by a single Gaussian function. This means that, from a static and averaged view, the correlations/anti-correlations between substates in consecutive steps are leading to compensatory effects.

In addition to the backbone movements and h-bonds, each substate at the TpA step is modulated and stabilized by other factors, such as interactions with ions and stacking between consecutive bases. For CpG, a relatively simple mechanism was found where the entrance of Na⁺/K⁺ inside the minor groove triggered and stabilized the low twist state and hence BII (18). For TpA, the mechanism is much more complex, since it involves a combination of shift/slide/twist substates and the movements of K⁺ from the major groove of CpT to the major groove of ApG, when going from HPN (high twist/positive slide/negative shift) to HPP (high twist/positive slide/positive shift) (Sup-

plementary Figure S6). A depletion of cations inside both grooves for the whole tetranucleotide was observed when moving to the LNZ substate (low twist/negative slide/zero shift). All the sequences studied share the same redistribution of K⁺ when moving between the substates, but the sequence-specific populations of each substate lead to different overall ion distributions when changing the next-to-nearest neighbour's context (Supplementary Figure S7). Finally, we found that at the TpA step, the stacking strength on either strand increased significantly when shift moves toward the minor groove at high twist and positive slide, an interaction that further stabilizes the BII state at the 3' junction (Supplementary Figure S8).

Structural information travels beyond next-to-nearest neighbours

Sequences studied here cover all the next-to-nearest neighbours' space with some redundancy that allowed us to check for some remote effects beyond this level. As noted above, such effects are clearly visible already in Figure 5, where sequences containing the same hexanucleotide sequence appear in two very different branches of the dendrogram, indicating the tuning of hexanucleotide preferences by more remote effects. Analysis of the different octanucleotide en-

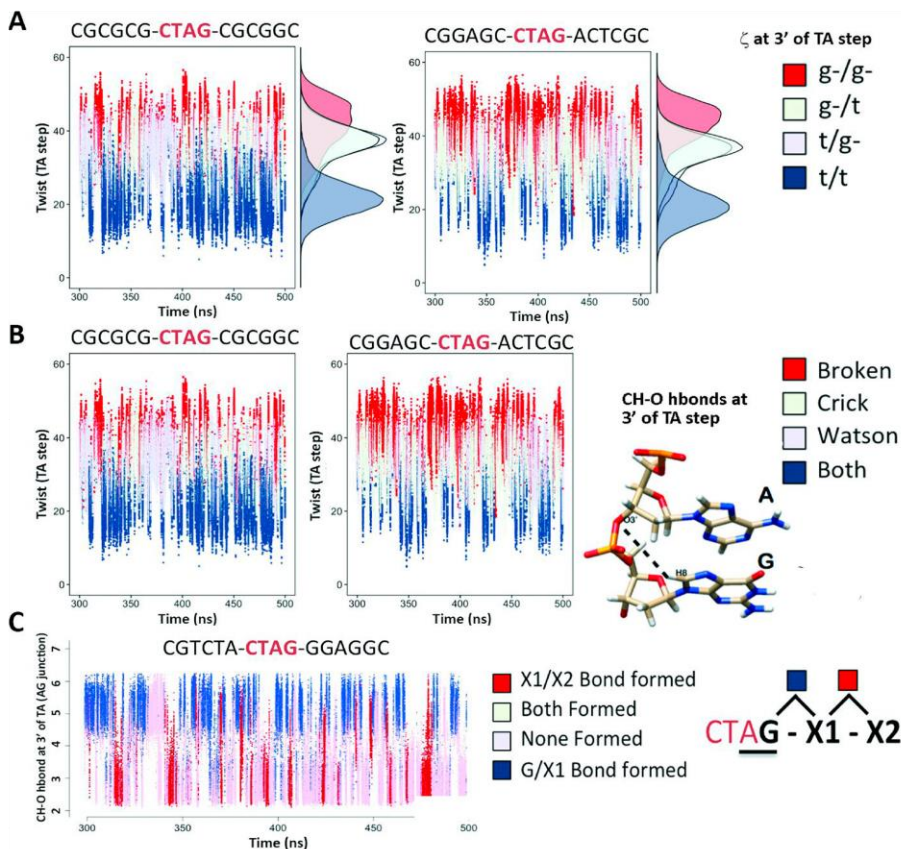


Figure 6. Concerted movements along the backbone and the bases explain the flow of structural information from the central TpA step and beyond next-to-nearest neighbours. (A) Correlation between twist and the BI/BII population (reduced to the ζ torsion at the 3'-side of TpA) at the TpA junction. (B) Correlation between twist at TpA and the CH-O h-bond formed at the ApG junction (bps + 1). (C) Correlation between the CH-O h-bond at the ApG junction with the CH-O h-bond at bps+1 (hexanucleotide context) and bps+2 (octanucleotide context). Note that the CH-O h-bonds are always coupled to BII propensities, stabilizing the BII substrate.

vironments (R·R/Y·Y), (Y·R) and (R·Y) reveals the existence of a quite differential behaviour (see Figure 7). For example, the conformational substates of the central TpA step in YpCpTpApGpR sequences (Y·R) are fully defined at the next-to-nearest neighbours level, with remote effects being negligible: all (Y·R) hexanucleotides appear in the same cluster in the dendrogram of Figure 5, and they display consistent distributions in all multi-modal helical parameters (shift has two main populations at ± 2 Å, with the zero shift state being less favoured). Slide and Twist are, as a consequence, pushed towards higher values. This makes sense, considering that, irrespective of the octanucleotide level base, when ApG is followed by an R base on both strands, the junction at ApG will be pushed out of the BII state. This frustration of high BII propensity of

two adjacent bps (a direct consequence of the crankshaft effect) will result in an overall higher BI population at ApG, which corresponds to the high twist, positive slide and negative/positive shift equilibrium at TpA. On the contrary, R·Y hexanucleotides (RpCpTpApGpY sequences) have two very distinct behaviours depending on the next flanking base: Central TpA steps in RpRpCpTpApGpYpY (RR·YY) octanucleotides tend to populate zero shift states and have equal populations of high/low twist as well as of negative/positive slide. On the contrary, TpA in YR·YR octanucleotide contexts have a strong preference for positive shift and rarely visit low twist or negative slide. Inspection of the trajectories suggests that this is probably due to a domino effect of h-bond proclivity so that depending on the base pairs flanking the R·Y hexanucleotide there is ei-

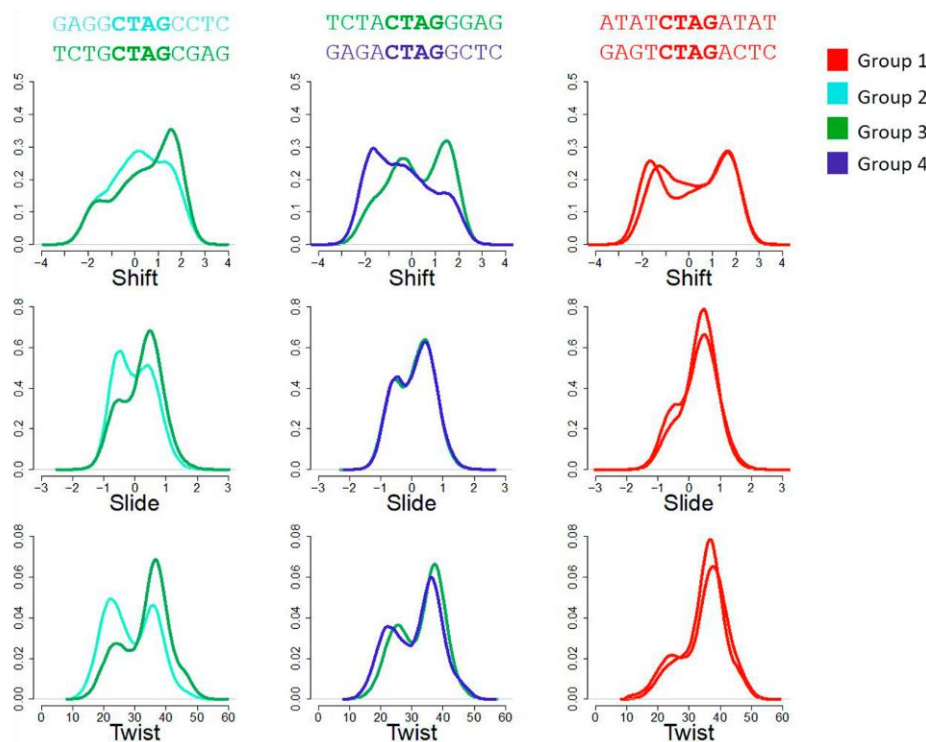
4426 *Nucleic Acids Research*, 2019, Vol. 47, No. 9

Figure 7. Normalized frequencies of shift, slide and twist at the central TpA step for three pairs of selected sequences showing non-negligible effects beyond next-to-nearest neighbours. The colours used are related to the groups found in the clustering analysis.

ther an equally strong preference towards BII at ApG on the two strands, or the Watson strand BII state is favoured over the Crick, which is necessarily compensated by shifting the bases towards the major groove. Finally, remote sequence effects are present just in a few cases for R \cdot R/Y \cdot Y hexanucleotides and lead to a change in shift from the minor to the major groove, maintaining similar distributions of twist and slide (Figure 7). In summary, our results suggest that CTAG is one of the few tetranucleotides (amongst the unique 136) where next-to-nearest neighbours and beyond effects are observed, while in general, nearest neighbour models can accurately explain by ‘concatenation of tetranucleotides’ the described remote effects in longer sequences.

Data mining of structural databases and genomic implications

We analysed the structures of DNA obtained experimentally (X-ray and NMR) and stored in the Protein Data Bank that contained the CTAG tetranucleotide sequence in order to validate our results. Only 106 occurrences of CTAG in naked DNA structures were found (some with small ligands or metal ions), and 160 occurrences in structures of

protein–DNA complexes. Moreover, only a fraction of the tetranucleotide sequence space is covered (next-to-nearest neighbours), and barely any of the hexanucleotide context (octanucleotides of the type XpXpCpTpApGpXpX, where X = C, T, A, G) is found (Supplementary Table S3). This scarcity of data clearly limits the generality of the conclusions that could be derived from the data mining of the PDB, although a BIC analysis of the experimental structural parameters of TpA steps flanked by 5′C-3′G at least confirms that multi-modality is not a force field artefact (Supplementary Figure S9). PDB structures containing the CTAG tetranucleotide have values for the shift, slide, roll and twist helical parameters that cover the multi-modal distributions obtained in our trajectories, confirming our claims on the bimodal nature of slide and twist, with peaks in the distributions that fit well to our results (see Figure 8 and Supplementary Figure S9). For shift, TpA steps distribution displays peaks 2 Å towards both the minor or major groove in several protein-bound DNA structures, but the data on naked DNA seem to be insufficient to cover these deformations: there is a small peak at +2 Å, but highly underestimated compared to our results. Finally, roll has a

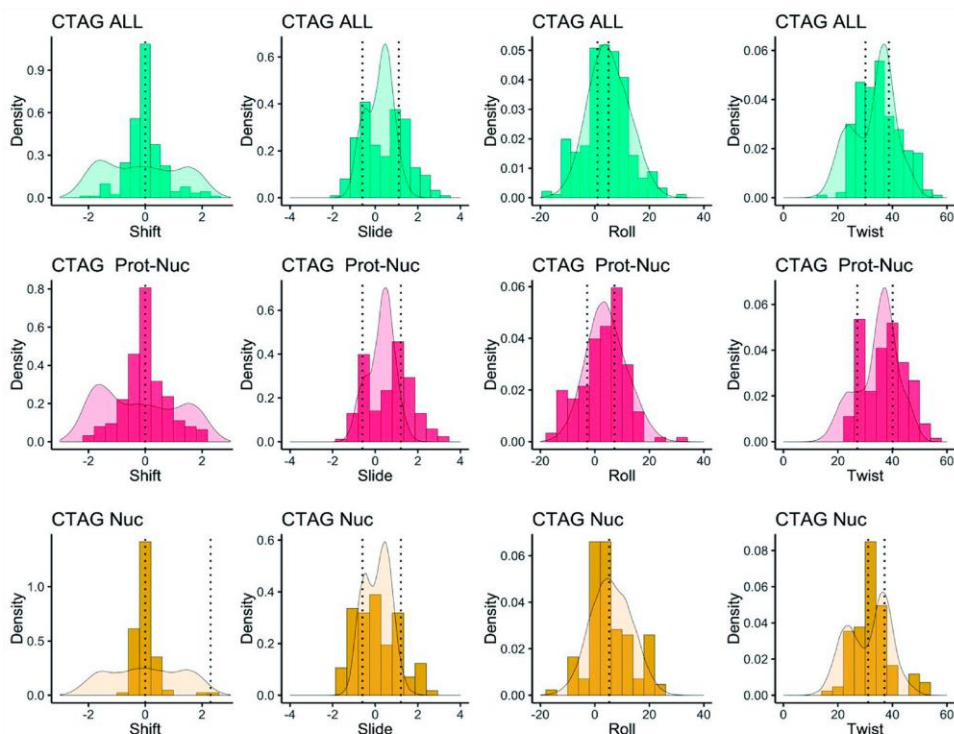


Figure 8. Normalized frequencies of shift, slide, roll and twist from MD meta-trajectory of representative hexanucleotides (G-C for free DNA; A-G, G-A, A-T and T-A for protein-bound DNA and their combination for all DNA structures) compared to those obtained from the data mining of the PDB for all structures containing DNA (first row), for Protein-DNA complexes (second row) and for isolated DNA structures (third row). The mean values of the BIC components of the experimental helical parameters data are shown as vertical dotted lines in each case.

broad distribution, similar to what we obtain from MD simulations, being bi-normal, but unimodal.

All analyses performed in this work suggests that CTAG has really unique physical properties, which should provide the genome with a point of high flexibility and polymorphism. Very remarkably, CTAG is one of the lowest populated tetranucleotides in the analysed species (see Figure 9) appearing mainly on intergenic regions and very rarely on genes (Supplementary Figure S10). We further highlighted this by analysing comparatively, with and without including exons, all the tetranucleotides containing the trinucleotide TpApG (XTAG or TAGX, where X could be A, C, G or T), which is known as the amber stop codon. Our results still confirm the low rate of the CTAG tetranucleotide, even removing the TpApG stop codon (Supplementary Figure S11). Interestingly, this infrequent CTAG tetranucleotide is well conserved, which suggest that (i) despite being far from coding regions they are important for the functionality of the cell, or alternatively, (ii) they are easily accessible to the mismatch repairing machinery, avoiding the stabilization of mutations. The same conclusion can be reached from

the analysis of cancer genomic data, which show that again CTAG is very rarely mutated in cancer (Supplementary Figure S12). The unusual physical properties of the CTAG tetranucleotide matches its unusual prevalence and distribution in the genome and its extreme resilience to somatic (cancer) mutations. It is tempting to believe that cell takes advantage of the unusual properties of CTAG as points of high flexibility that might help to fold chromatin.

CONCLUSIONS

We present here an in-depth study of one of the most 'structurally speaking' polymorphic tetranucleotides found in B-DNA. The complete helical space of the CTAG tetranucleotide has been analysed by means of extensive molecular dynamics simulations and by data mining the Protein Data Bank, confirming its highly polymorphic behaviour at several helical parameters: shift, slide, twist and BI/BI1. This confers to CTAG the possibility to exist in several different substates, being particularly flexible. We present here clear evidence that the type of substate displayed by CTAG in a given sequence context, and in conse-

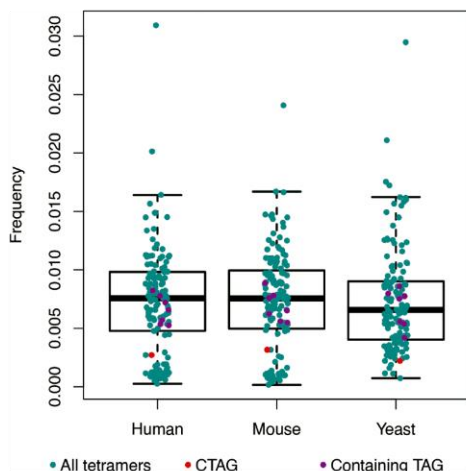


Figure 9. Frequency of each possible tetranucleotide in three different genomes. CTAG is marked in red, tetranucleotides containing TpApG (amber stop codon) are marked in violet and the rest are depicted in cyan.

quence its dynamics, is sequence dependent, and fine-tuned by next-to-nearest neighbours and beyond. Based on the concerted and correlated movements of bases and backbone torsions for the described multi-modal degrees of freedom, and driven by the mechanical limitations imposed by DNA's crankshaft motions, we were able to found a possible explanation on how structural information can travel almost half helical turn away from the central TpA step. This remote structural 'connection' allows the TpA step to 'feel' its sequence environment beyond the next-to-nearest neighbours, and eventually adopts a different substate if needed. Moreover, we found that previously described unconventional intra-molecular hydrogen bonds of the type C8H8-O3' and C6H6-O3' that link the movements of the bases with the torsions in the backbone, could be used as descriptors of such correlated motions. Finally, we established that although this highly flexible tetranucleotide is extremely under-represented in several genomes along the animal Kingdom, being mostly present in intergenic sequences, it has been preserved with a low rate of mutation implying a possible physical role for CTAG at the genomic level.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

M.O. is an ICREA (Institutió Catalana de Recerca i Estudis Avançats) academia researcher. P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher.

Author contributions: The sequence library was designed by P.D.D. and A.B. Simulations were performed by A.B., with the assistance of J.W. and P.D.D. Analysis of the simulations was designed and performed by A.B., with all authors involved in assessing results and further discussions. D.B. did the genome-wide analysis, and A.H. helped with the data mining of PDB structures. M.O. and P.D.D. integrated all the results, discussed the analysis and wrote the manuscript with contributions from all the co-authors. The original idea of the project came from P.D.D. and M.O.

FUNDING

Spanish Ministry of Science [BFU2014-61670-EXP, BFU2014-52864-R]; Catalan SGR, Instituto Nacional de Bioinformática; European Research Council (ERC SimDNA); European Union's Horizon 2020 Research and Innovation Program [676556]; Biomolecular and Bioinformatics Resources Platform (ISCH PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) (to M.O.); MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). Funding for open access charge: European Union's Horizon 2020 Research and Innovation Program [676556].

Conflict of interest statement. None declared.

REFERENCES

- Wilkins, M.H.F., Stokes, A.R. and Wilson, H.R. (1953) Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature*, **171**, 738–740.
- Franklin, R.E. and Gosling, R.G. (1953) Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–741.
- Lucas, A.A., Lambin, P., Mairesse, R. and Mathot, M. (1999) Revealing the backbone structure of B-DNA from laser optical simulations of its X-ray diffraction diagram. *J. Chem. Educ.*, **76**, 378.
- Kypr, J., Kejnovská, I., Rencuk, D. and Vorlicková, M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.*, **37**, 1713–1725.
- Kato, M. (1999) Structural bistability of repetitive DNA elements featuring CA/TG dinucleotide steps and mode of evolution of satellite DNA. *Eur. J. Biochem.*, **265**, 204–209.
- Kielkopf, C.L., Ding, S., Kuhn, P. and Rees, D.C. (2000) Conformational flexibility of B-DNA at 0.74 Å resolution: d(CCAGTACTGG)2. *J. Mol. Biol.*, **296**, 787–801.
- Maehigashi, T., Hsiao, C., Kruger Woods, K., Moulai, T., Hud, N.V. and Dean Williams, L. (2012) B-DNA structure is intrinsically polymorphic: even at the level of base pair positions. *Nucleic Acids Res.*, **40**, 3714–3722.
- Monchaud, D., Allain, C., Bertrand, H., Smargiasso, N., Rosu, F., Gabelica, V., De Cian, A., Mergny, J.-L. and Teulade-Fichou, M.-P. (2008) Ligands playing musical chairs with G-quadruplex DNA: A rapid and simple displacement assay for identifying selective G-quadruplex binders. *Biochimie*, **90**, 1207–1223.
- Radhakrishnan, I. and Patel, D.J. (1994) DNA Triplexes: Solution structures, hydration sites, energetics, interactions, and function. *Biochemistry*, **33**, 11405–11416.
- Kaushik, M., Kaushik, S., Bansal, A., Saxena, S. and Kukreti, S. (2011) Structural diversity and specific recognition of four stranded G-quadruplex DNA. *Curr. Mol. Med.*, **11**, 744–769.
- Dai, J., Carver, M. and Yang, D. (2008) Polymorphism of human telomeric quadruplex structures. *Biochimie*, **90**, 1172–1183.
- Dans, P.D., Pérez, A., Faustino, I., Lavery, R. and Orozco, M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
- Dans, P.D., Danilâne, L., Ivani, I., Dršata, T., Lankaš, F., Hospital, A., Walther, J., Pujagut, R.I., Battistini, F., Gelpi, J.L. et al. (2016)

- Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.*, **44**, 4052–4066.
14. Imeddourene, A. Ben, Xu, X., Zargarian, L., Oguey, C., Foloppe, N., Mauffret, O. and Hartmann, B. (2016) The intrinsic mechanics of B-DNA in solution characterized by NMR. *Nucleic Acids Res.*, **44**, 3432–3447.
 15. Ben Imeddourene, A., Elbahnsi, A., Guéroult, M., Oguey, C., Foloppe, N. and Hartmann, B. (2015) Simulations meet experiment to reveal new insights into DNA intrinsic mechanics. *PLOS Comput. Biol.*, **11**, e1004631.
 16. Tian, Y., Kayatta, M., Shultis, K., Gonzalez, A., Mueller, L.J. and Hatcher, M.E. (2009) ³¹P NMR investigation of backbone dynamics in DNA binding sites¹. *J. Phys. Chem. B*, **113**, 2596–2603.
 17. Zgarbová, M., Jurečka, P., Lankaš, F., Cheatham, T.E., Šponer, J. and Otyepka, M. (2017) Influence of BII backbone substates on DNA Twist: A unified view and comparison of simulation and experiment for all 136 distinct tetranucleotide sequences. *J. Chem. Inf. Model.*, **57**, 275–287.
 18. Dans, P.D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R. and Orozco, M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
 19. Balaceanu, A., Pérez, A., Dans, P.D. and Orozco, M. (2018) Allostery and signal transfer in DNA. *Nucleic Acids Res.*, **46**, 7554–7565.
 20. Calladine, C.R., Drew, H.R., Luisi, B.F. and Travers, A.A. (2004) *Understanding DNA: The molecule and how it works*. Elsevier Academic Press, London and San Diego.
 21. Dickerson, R.E. and Klug, A. (1983) Base sequence and helix structure variation in B and A DNA. *J. Mol. Biol.*, **166**, 419–441.
 22. Fratini, A. V., Kopka, M.L., Drew, H.R. and Dickerson, R.E. (1982) Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *J. Biol. Chem.*, **257**, 14686–14707.
 23. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
 24. Cheatham, T.E., Cieplak, P. and Kollman, P.A. (1999) A modified version of the Cornell et al. Force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
 25. Pérez, A., Marchán, I., Svozil, D., Šponer, J., Cheatham, T.E., Laughton, C.A., Orozco, M. and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
 26. Dršata, T. and Lankaš, F. (2015) Multiscale modelling of DNA mechanics. *J. Phys. Condens. Matter*, **27**, 323102.
 27. Dršata, T., Pérez, A., Orozco, M., Morozov, A. V., Šponer, J. and Lankaš, F. (2013) Structure, stiffness and substates of the Dickerson–Drew dodecamer. *J. Chem. Theory Comput.*, **9**, 707–721.
 28. Ivani, I., Dans, P.D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A. et al. (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
 29. Dans, P.D., Ivani, I., Hospital, A., Portella, G., González, C. and Orozco, M. (2017) How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.*, **45**, 4217–4230.
 30. Jiang, W., Phillips, J.C., Huang, L., Fajer, M., Meng, Y., Gumbart, J.C., Luo, Y., Schulten, K. and Roux, B. (2014) Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. *Comput. Phys. Commun.*, **185**, 908–916.
 31. Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. and Walker, R.C. (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.
 32. Lee, J., Cheng, X., Swails, J.M., Yeom, M.S., Eastman, P.K., Lemkul, A., Wei, S., Buckner, J., Jeong, J.C., Qi, Y. et al. (2016) CHARMM-GUI Input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.*, **12**, 405–413.
 33. Páll, S., Abraham, M.J., Kutzner, C., Hess, B. and Lindahl, E. (2015) *Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS*. Springer, Cham, Stockholm, pp. 3–27.
 34. Beveridge, D.L., Barreiro, G., Suzie Byun, K., Case, D.A., Cheatham, T.E., Dixit, S.B., Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H. et al. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA Oligonucleotides. I. research design and results on d(CpG) steps. *Biophys. J.*, **87**, 3799–3813.
 35. Dixit, S.B., Beveridge, D.L., Case, D.A., Cheatham, T.E., Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H., Osman, R., Sklenar, H. et al. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA Oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**, 3721–3740.
 36. Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dixit, S., Jayaram, B., Lankas, F., Laughton, C. et al. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, **38**, 299–313.
 37. Pasi, M., Maddocks, J.H., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dans, P.D., Jayaram, B., Lankas, F., Laughton, C. et al. (2014) μABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
 38. Pasi, M., Maddocks, J.H. and Lavery, R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–2423.
 39. Balaceanu, A., Pasi, M., Dans, P.D., Hospital, A., Lavery, R. and Orozco, M. (2017) The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**, 21–28.
 40. Hospital, A., Andrio, P., Cugnasco, C., Codo, L., Becerra, Y., Dans, P.D., Battistini, F., Torres, J., Goñi, R., Orozco, M. et al. (2016) BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
 41. Yuan, H., Quintana, J. and Dickerson, R.E. (1992) Alternative structures for alternating poly(dA-dT) tracts: the structure of the B-DNA decamer C-G-A-T-A-T-A-T-A-T-C-G. *Biochemistry*, **31**, 8009–8021.
 42. Mack, D.R., Chiu, T.K. and Dickerson, R.E. (2001) Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. *J. Mol. Biol.*, **312**, 1037–1049.
 43. Case, D.A., Betz, R.M., Cerutti, D., Cheatham, T.E. III, Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N. et al. (2016) AMBER 2016.
 44. Berendsen, H.J.C., Grigera, J.R., Straatsma, T.P., Grigera, J.R., Straatsma, T.P., Berendsen, H., Grigera, J., Straatsma, T., Grijera, J., Berendsen, H.J.C. et al. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
 45. Smith, D.E. and Dang, L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
 46. Dang, L.X. (1995) Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-Crown-6 Ether: A molecular dynamics study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
 47. Dang, L.X. and Kollman, P.A. (1995) Free energy of association of the K⁺:18-Crown-6 complex in Water: A new molecular dynamics study. *J. Phys. Chem.*, **99**, 55–58.
 48. Darden, T., York, D. and Pedersen, L. (1993) Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
 49. Ryckaert, J.-P., Cicotti, G. and Berendsen, H.J. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
 50. Hospital, A., Faustino, I., Collepardo-Guevara, R., González, C., Gelpi, J.L. and Orozco, M. (2013) NAFlex: a web server for the study of nucleic acid flexibility. *Nucleic Acids Res.*, **41**, W47–W55.
 51. Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
 52. Schwarz, G. (1978) Estimating the dimension of a Model. *Ann. Stat.*, **6**, 461–464.
 53. Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
 54. Schilling, M.F., Watkins, A.E. and Watkins, W. (2002) Is human height bimodal? *Am. Stat.*, **56**, 223–229.

4430 *Nucleic Acids Research*, 2019, Vol. 47, No. 9

55. Ghosh, A. and Bansal, M. (2003) A glossary of DNA structures from A to Z. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **59**, 620–626.
56. Jolliffe, I.T. (1986) *Principal Component Analysis*. Springer-Verlag, NY.
57. Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–441.
58. Dršata, T. and Lankaš, F. (2013) Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 355–363.
59. Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
60. Murtagh, F. and Legendre, P. (2014) Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's Criterion? *J. Classif.*, **31**, 274–295.
61. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of Lineage-determining transcription factors Prime cis-Regulatory elements required for macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.
62. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.-L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
63. Zerbino, D.R., Achuthan, P., Akanni, W., Mader, M.R., Barrell, D., Bhaki, J., Billis, K., Cummins, C., Gall, A., Giron, C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

3.2 Sequence selective protein-DNA recognition

In the previous section, the polymorphic nature of a specific tetramer sequence was explored. Next, we studied a broader set of DNA structures to understand whether the DNA physical properties are enough to explain the structural difference between the naked DNA sequences and the bioactive conformation found in protein-DNA complexes taken from the PDB. The study of the structures of the protein-bound DNA available revealed the prevalence of a shape readout model on protein recognition of DNA binding sites. Here, two mechanisms can lead to the DNA to adopt the bioactive conformation:

- Conformational selection when the structural changes occur spontaneously in the absence of the protein. Then, the deformation energy required to adopt the protein-bound conformation is within the energy range sampled by the free DNA (thermal energy fluctuations).
- Induced fit when the changes occur after the binding event. In this case the DNA is highly distorted by the protein and consequentially the deformation energy required is high.

In this work, we selected structures from the PDB containing protein-DNA complexes, applying a set of filters to discard those comprising single-stranded or extremely distorted B-DNA conformations (for instance including opened base-pairs), and investigated the evidence towards conformational selection or induced fit in these complexes using MD simulations and essential dynamics analyses. The helical parameters were collected from the MD simulations of the naked DNA and a statistical test was computed to evaluate whether the values explored by each physical descriptor contained those from the experimentally determined structures. We used Hotelling's multivariate statistical test for each helical movement, computing

$$F = \frac{n-m}{m} (\mu - \bar{x})^t S^{-1} (\mu - \bar{x}) \quad (3.1)$$

where n is the number of frames in the MD simulation, m is the number of base pair steps in each sequence, μ is the vector containing the observed values of the

helical parameter in the PDB structure, \bar{x} is the vector containing the average values along time of the helical parameters and S the covariance matrix of these values. A significant F value (that is $F > F_{1-\alpha;m,n-m}$ at $1 - \alpha = 95\%$, where $F_{1-\alpha;m,n-m}$ is the $1 - \alpha$ quantile of an F distribution with $m, n - m$ degrees of freedom) indicates that the bound conformation is not sampled by the naked DNA trajectory. With this analysis we found that in a large proportion of the cases, the helical motions required for the bound conformation are sampled by naked DNA.

Then, we characterized those structures having a significant Hotelling's statistic, i.e. the conformation in the protein-DNA complex is not sampled by the naked DNA simulations. Detailed statistical analysis of each base pair step helical parameter revealed the positions that caused the global F value to be large, comparing the experimental value in the complex with the values sampled in the naked DNA simulations. When large distortions are observed in the protein-bound conformation, usually they are found at regions that directly interact with the bound protein, linked to changes in the backbone angles and to the tendency of the phosphates to approach cationic residues.

Additionally, computing the deformation energy associated with protein-DNA binding, we observed the prevalence of a conformational selection in a large proportion of the cases (71%) over a small percentage where the induced-fit was the major driver for the complex formation (11%) (the remaining 18% is in a zone where both processes might be occurring).

In summary, the statistical analysis of our trajectories supports the shape readout mechanism of protein-DNA binding. However, although the sequence dependent physical properties are important for adopting the required conformation in most of the complexes analyzed, they are not sufficient to explain the mechanism of protein binding in all the cases as specific sequence-reading may contribute to significant DNA-protein interaction.

Publication:

Federica Battistini, Adam Hospital, Diana Buitrago, Diego Gallego, Pablo D. Dans, Josep Lluís Gelpí, and Modesto Orozco. (2019). How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition. *Journal of Molecular Biology* S0022-2836(19)30451-6. <https://doi.org/10.1016/j.jmb.2019.07.021>.

Supplementary material for this article can be found in the **Annex II**.



How B-DNA Dynamics Decipher Sequence-Selective Protein Recognition

Federica Battistini¹, Adam Hospital¹, Diana Buitrago¹, Diego Gallego¹,
Pablo D. Dans¹, Josep Lluís Gelpí² and Modesto Orozco^{1,2}

¹ - Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

² - Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain

Correspondence to Modesto Orozco: Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. modesto.orozco@irbbarcelona.org

<https://doi.org/10.1016/j.jmb.2019.07.021>

Edited by Monika Fuxreiter

Abstract

The rules governing sequence-specific DNA–protein recognition are under a long-standing debate regarding the prevalence of *base versus shape* readout mechanisms to explain sequence specificity and of the *conformational selection versus induced fit* binding paradigms to explain binding-related conformational changes in DNA. Using a combination of atomistic simulations on a subset of representative sequences and mesoscopic simulations at the protein–DNA interactome level, we demonstrate the prevalence of the *shape readout* model in determining sequence-specificity and of the *conformational selection* paradigm in defining the general mechanism for binding-related conformational changes in DNA. Our results suggest that the DNA uses a double mechanism to adapt its structure to the protein: it moves along the easiest deformation modes to approach the bioactive conformation, while final adjustments require localized rearrangements at the base-pair step and backbone level. Our study highlights the large impact of B-DNA dynamics in modulating DNA–protein binding.

© 2019 Elsevier Ltd. All rights reserved.

Introduction

DNA–protein recognition, an essential step in gene regulation, depends on both the accessibility of the DNA and its intrinsic affinity for the protein. Accessibility is related to the chromatin fold and to the presence of competing proteins, while affinity is determined by the formation of protein–DNA contacts and by the cost of deforming the DNA duplex from the naked to the bound bioactive conformation. Two extreme situations can be envisioned in DNA–protein binding: one where the complex formation follows a base readout mechanism in which specific DNA–protein contacts determine sequence specificity, and another one where the binding follows an *shape readout* model; that is, DNA deformability properties explain sequence-specific binding [1]. *Shape* (indirect) readout describes protein–DNA recognition mechanisms that depend on the ability of a DNA sequence to adopt a conformation that facilitates its binding to the protein or that intrinsically

has the matching conformation for the protein binding. Protein–DNA *shape* recognition involves the formation of specific binding sites for positively charged amino acids, ARG/LYS, indirect contacts with phosphates, some direct hydrogen bonds established with DNA bases and interactions mediated through water molecules. Those interactions depend on the different solvation, above all, upon binding, the release of water molecules from the protein–DNA interface provide a favorable entropic contribution, and it is important for selectivity [2–4]. Very often, protein binding leads to a conformational change in DNA, and again two different models can be proposed to explain the connection between structural flexibility and binding: *conformational selection* and *induced fit*. The recognition modes contribute to the overall protein–ligand binding mechanism that couples conformational selection and conformational changes, which depend on the ligand, in this case DNA, and protein properties and on multiple conditions, including the interactions

between the biomolecules, their concentrations [5] and the rate of the conformational transition [6]. Within the conformational selection paradigm, the deformation energy required to move the DNA from naked to bioactive conformation is small, typically within the DNA thermal fluctuation, and it is then sampled spontaneously in the “unbound” state. On the contrary, according to the induced fit model the DNA deformation energy required for binding is large, hampering the spontaneous population of the bound state from the naked B-DNA dynamics.

In the last decades, many experimental and computational studies analyzed the specificity of the protein–DNA binding to better understand their recognition [7–15]. Thus, databases have been built and software has been developed to study the interactions, affinity and selectivity in protein–DNA binding [7]. Databases store, for example, preferred DNA binding sites for a large number of proteins as determined by SELEX-seq/HT-SELEX, microarrays, chromatin immunoprecipitation and others [8–17]. Such sequence-based information is combined with the structural analysis of known protein–DNA complexes to derive interaction rules, which are implemented in a variety of statistical methods [7,18–22]. Alternatively, *ab initio* approaches to the study of protein–DNA interactions are based on the use of energy-based *in silico* methods, which use protein–DNA direct interaction terms [21], and deformation energies derived from DNA properties [22] to recognize binding sites through structural signals.

Despite the variety of experimental and computational studies on DNA–protein binding, the relative importance of base *versus* shape readout is unclear, and no consensus exists on the prevalence of induced fit or conformational selection paradigms [1]. Certainly, part of the problem is due to the discrepancy existing in the experimental information available, as data obtained from HT-SELEX [17,23], footprinting [24], protein binding microarrays [8,25] or ChIP-chip/ChIP-seq experiments [12,26,27] depend on the experimental technique and conditions making statistical methods noisy and often over-trained to reproduce just one type of data. For this reason, experimentally trained statistical methods should be complemented with approaches based on the calculation of interaction and deformation energies by means of physical models, which are not influenced by the noise of high-throughput experimental data [28–38].

In this article, we present an *in silico* analysis of the role of DNA conformational flexibility in the formation of protein–DNA complexes. The systematic evaluation of the conformational changes of physiological DNA associated with protein binding was performed using molecular dynamics simulations, with the newly refined parmbsc1 force field, which allowed performing analyses of DNA structure and flexibility with accuracy similar to that of current experimental

techniques [39–43], and mesoscopic simulations, focusing on DNA sequence preferences. Results presented here provide convincing evidence for the impact of the shape readout on the DNA–protein interactome, and for the prevalence of conformational selection mechanism in defining binding-related conformational change in DNA, at least in those cases where the protein does not have a clear disruptive effect on the DNA structure. Our results suggest that DNA adapts to the presence of the interacting protein following a dual mechanism: global movements are facilitated as coded in the essential dynamics of the duplex, while local rearrangements are related to displacements at the base-pair step level and are coupled to complex backbone rearrangements. In this analysis, we took into account that the torsion angles from experimental data (NMR and x-ray) are difficult to determined and are not completely captured and validated (for a discussion on experimental backbone torsion angles reliability, see Refs. [44, 45]). However, results presented here show how sequence-dependent B-DNA dynamics are a key player in modulating DNA–protein recognition and that dynamics of isolated DNA in physiological conditions is important in determining DNA–protein interaction, independently of the specific Protein–DNA binding motif.

Results and Discussion

MD simulations of the 50 naked DNA sequences (see [Materials and Methods](#), [Fig. 1](#)) provided stable trajectories without any remarkable distortion after 500 ns of simulation time. The origin of the starting structure (canonical B-form or bound state) is not relevant (see Supplementary Fig. S2) supporting the idea that simulations are sampling equilibrium conformations without memory of the initial structure. The conformational space sampled in the trajectories agrees very well with the one expected for B-DNA duplexes [39], leading to a set of structures that lost memory of the initial experimental ones (x-ray or NMR) [39,40,46].

How is the intrinsic geometry of the DNA modified by protein binding?

The comparison between the experimental DNA structure and the conformational space sampled by the naked DNA in the MD simulations using Hoteling's statistics (see [Materials and Methods](#)) revealed that in general the bound DNA structure falls within naked DNA conformational space (red circle in [Fig. 2](#) for rise and roll and Supplementary Fig. S3 for the remaining bp parameters). In the few cases, where DNA ensembles show local differences from the bound DNA structure, we checked meticulously for potential uncertainties in the

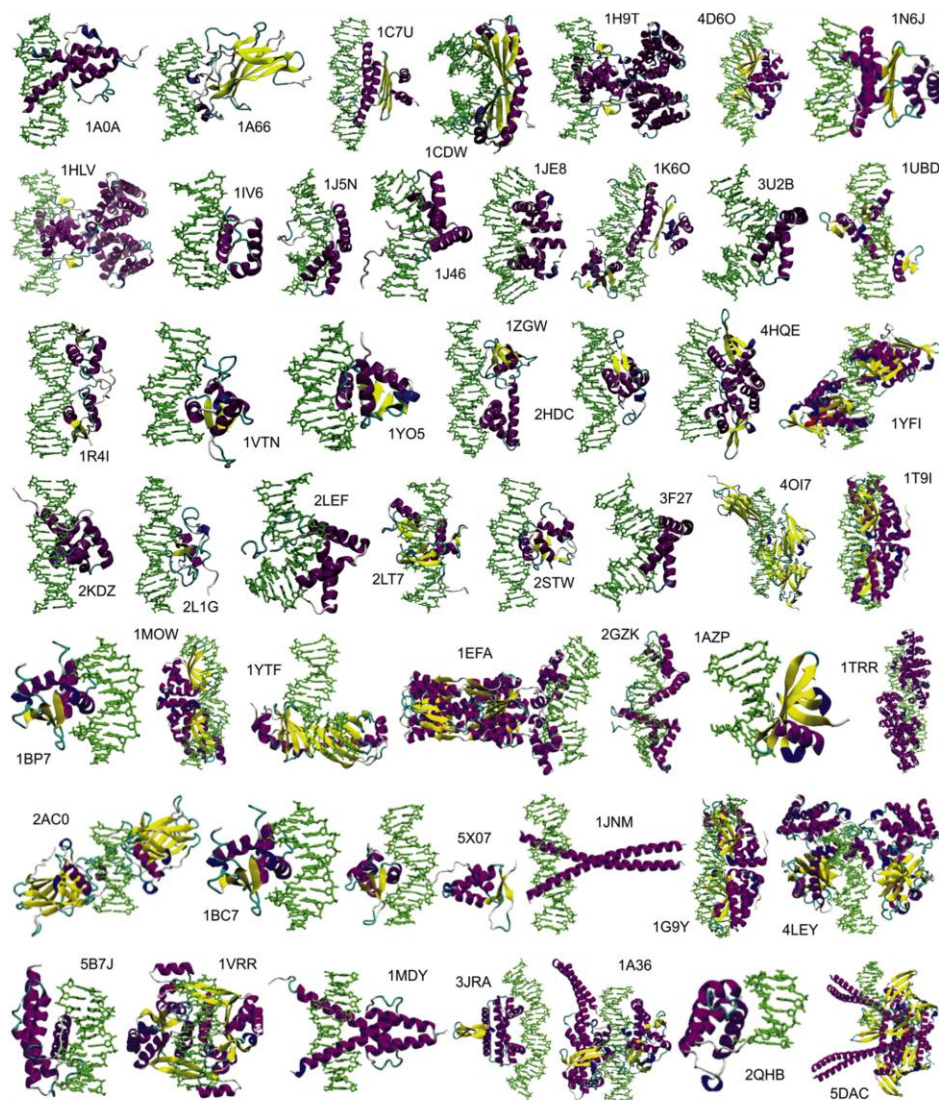


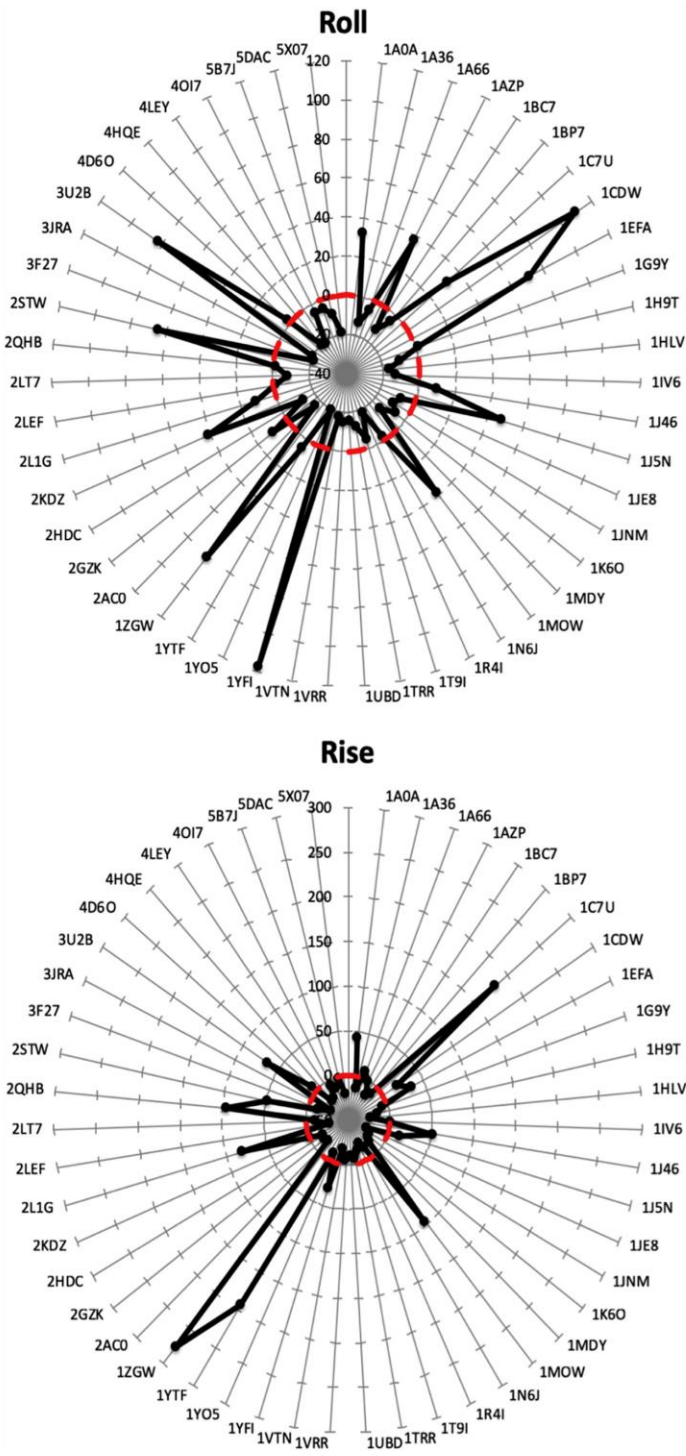
Fig. 1. Representation of the protein–DNA complexes summarized with details in Table S1 [7]. The PDB IDs are indicated.

reported experimental structure. We found three NMR structures where doubts may exist regarding certain structural details. For example, 1C7U shows highly unusual rise, slide and roll values (see Supplementary Fig. S4), in regions away from the protein, signaling potential artifacts in the refinement leading to abrupt and compensatory helical profiles [40]. 1ZGW shows an unusual rise profile at the duplex termini ($d(A_3A_4)$ and $d(A_{15}A_{16})$); the large rise in the latter may be

explained by the partial intercalation of Phe₁₁₄, but the large and unusual rise (around 5 Å) at the other base-pair step is very difficult to explain as there are no interacting protein residues in the vicinity (see Supplementary Fig. S5). Finally, 2STW shows further unusual rise values (see Supplementary Fig. S6) at the central $d(T_7T_8)$ step, which can be explained by the presence of the partial intercalation of Tyr₈₅, and at the termini ($d(C_2G_3)$ and $d(C_{15}G_{16})$) where the high rise is suspicious as

3848

Sequence-Selective Protein Recognition



it is not justified by any protein–DNA interactions. Complexes 1T9I, 2KDZ, 2L1G, 1MOW, 1YTF, 2QHB, 1YFI, 1AZP, 1A0A, 1CDW, 3F27 and 3U2B show some unusual values in helical parameters of the DNA (Fig. 2), which could, however, be explained by direct contacts with the protein. For example, partial intercalation explains the large kink in the last 3 structures (1CDW, 3F27 and 3U2B), while strong salt-bridge contacts of DNA backbone with cationic residues of the protein could explain unusual roll and rise values in 1A0A (see Fig. 3 and Supplementary Figs. S7–9).

What is the backbone conformation required for protein binding?

In general, backbone angles in DNA–protein complexes remained in the conformational space sampled by naked DNA simulations, like the base-pair parameters (Fig. 2). However, when large alterations in helical coordinates are required, they are achieved by concerted changes in the backbone angles (α , β , χ , ϵ , γ , phase and ζ) [47,48]. Kinks, highly bent base-pair steps, are linked to significant alterations in sugar pucker, which are rare in unperturbed DNA and can be coupled to other backbone changes. It seems that coordinated movements of α/γ and, to a lesser extent, ϵ/ζ are frequent in protein-distorted DNA and are related to the tendency of the phosphates to approach cationic residues in the interacting protein. Changes use to be localized at regions of direct contact with the protein. An example is given in Fig. 3 that shows a detailed analysis of Protein Data Bank (PDB) ID 1A0A, where distortions in both helical parameters and backbone angles are visible at regions interacting with protein. In particular, base-pair steps with high roll (CC and CG) are correlated with unusual α/γ angle and are in contact with protein residues ARG, LYS and GLU (Fig. 3). We also detected correlations between distorted base-pair step parameters and unusual backbone values for α/γ and phase (in kinked structures also χ and β) angles where the protein residues are in contact with the DNA in particular for the structures PDB ID 1CDW, 3U2B and 3F27 (Supplementary Figs. S7–S9). Overall, our analysis conclude that unbound DNA backbone is rather flexible under physiological conditions, and there are few cases where unusual conformations of the backbone, not present in the naked ensemble, are required for adopting the bioactive conformation.

What is the energy cost of deforming helical coordinates for binding?

The thermal energy fluctuation calculated for naked DNA along the MD, amounts to around 2.5 ± 0.1 kcal/mol bp (see Materials and Methods) and, accordingly, when the energy cost of achieving the bound state is lower than this value, we can conclude that the bound state can be spontaneously sampled (being thermodynamically accessible at physiological conditions) by the naked DNA. Inside this energetic range, the DNA–protein binding follows a behavior that falls into the conformational selection paradigm. In contrast, when distortion energy cost is larger than this value, we can conclude that the DNA needs external effector to change structure and adopt the bioactive conformation, leading to the induced fit mechanism (Fig. 4a). We considered a margin of twice the free DNA fluctuation energy as a twilight zone (red area in Fig. 4a, between 2.5 and 5.0 kcal/mol bp), where hypothetically both recognition modes, conformational selection and induced fit, coexist [49].

Mesoscopic calculations (Fig. 4a) indicate that for 33 of the 50 complexes considered, the energy cost for the DNA to adapt to the bioactive conformation is within the free DNA fluctuation energy range and fall inside the defined blue area (Fig. 4a). That is, in most cases, binding follows the requirements of the conformational selection mechanism. The induced fit mechanism explains binding in 12 of the 50 complexes (white area in Fig. 4a), while the remaining 5 cases can be labeled as in the twilight zone (red area in Fig. 4a), where possibly both mechanisms contribute to the binding. Our results indicate that conformational selection seems to be at least twice more prevalent than induced fit in modulating DNA–protein binding in our set of representative DNA–protein complexes.

Are essential deformation modes coupled to protein-induced DNA deformation?

Large protein-induced conformational transitions in DNA (initial RMSD between naked and bound structure $\text{RMSD}_{\text{in}} > 5$ Å, value to delineate the boundary between large and small DNA distortion, defined by the average plus one standard deviation of the dots in the blue area in Fig. 4a) are possible thanks to good alignment between the transition vector (from the naked to the bound structure) and the essential deformation (ED) modes of the naked DNA (see Materials and Methods). Such an alignment is not a necessary for small protein-induced transitions

Fig. 2. Base-pair parameter confidence region profile. For each protein-bound DNA structure identified by their PDB ID, the axis represents the difference between the observed test statistic and the 95% critical value from the F distribution ($F - F_{(1-\alpha, m, n-m)}$). The value for each base-pair parameter, translation rise and rotational roll, can be inside (<0 , limit defined by red line) or outside the naked DNA conformational space (>0). See Materials and Methods and Supplementary Methods for discussion.

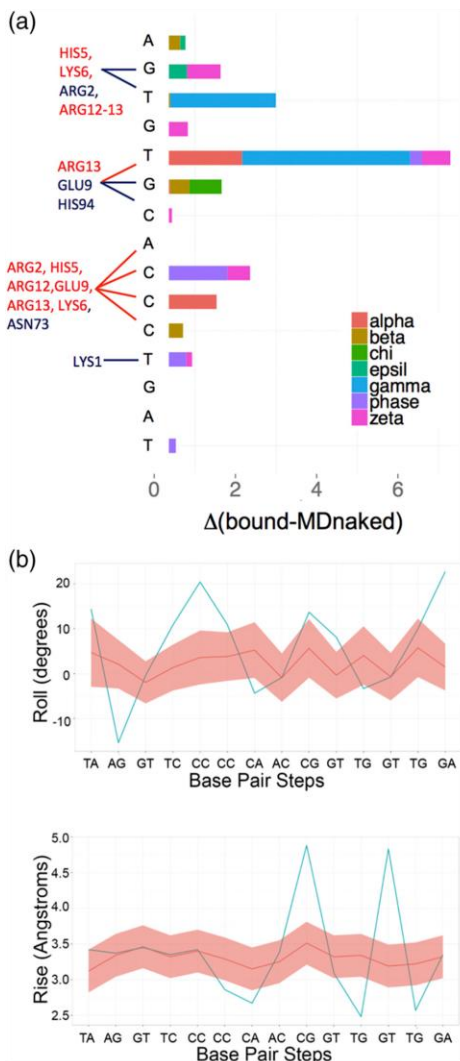


Fig. 3. Backbone and base-pair parameter analysis for the complex PDB ID 1A0A. (a) Analysis of the backbone angles ($\alpha, \beta, \chi, \epsilon, \gamma, \text{phase}, \zeta$) is shown. Backbone angle variation has been analyzed using the difference between the experimental protein-bound DNA angle values and the average MD simulated naked DNA values plus the standard deviation, divided by the standard deviation along the MD trajectory for each backbone angle ($\Delta(\text{bound-MDnaked})$). (b) Comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for base-pair step parameter roll and rise. The distortion given by the contact of the protein helices and coil residues (in red and blue, respectively in the image on top and named in the left panel) at the base level, extreme roll and rise values at steps CC, CG and GT, is correlated with deformation of the backbone angles in the backbone.

($\text{RMSD}_{\text{in}} < 5 \text{ \AA}$), where only local rearrangements are required. This is shown in the dependence of the RMSD_{in} and the squared overlap calculated between ED modes and the transition vector, as well as in the correlation between RMSD_{in} and the distance covered applying the ED modes of the naked DNA (see [Materials and Methods](#) and [Fig. 4b–c](#)). Our results strongly suggest that DNA adapts to protein shape following a dual mechanism. On one hand, global deformations happen along preferred deformation modes and bring the naked DNA structure close to that of the bioactive (protein-bound) conformation (around 3–4 Å in RMSD), independently of the original RMSD_{in} ([Fig. 4d](#)). On the other, small movements at the base-pair step level are required for the fine grain adjustments to reach for the perfect complementarity between the protein and the DNA.

We found that in general, physiological B-DNA is flexible enough to easily sample its bioactive conformation without the presence of the protein, supporting the prevalence of the conformational selection over the induced fit mechanism, at least for protein–DNA complexes where the protein does not break the Watson–Crick base pairing. In those cases where reaching the bioactive conformation implies mild distortions, they typically involve local rearrangements in the base-pair step geometry and small backbone changes. However, when the required distortion is large, DNA reaches the bioactive state by first moving along the low-energy ED modes, and finally by way of local rearrangements fine tuning DNA conformation sub-states.

What is the driving force for large protein-induced structural deformations?

As discussed above, many of the complexes studied here require structural distortions in the DNA that are easy to achieve from the naked ensemble, in agreement with the conformational selection model. There are, however, a few complexes for which conformational changes come at a high deformation cost (see [Fig. 4a](#)), and we were intrigued on the driving force of these distortions. A detailed analysis of these cases shows that the structural deformation induced by the protein results in changes in the electrostatic field of the DNA, which obey the need of DNA to accommodate to the protein interacting residues. Thus, upon binding, regions of the DNA facing apolar residues become less cation-philic, while negative potential is reinforced in those regions facing Arg/Lys-rich areas [see selected molecular interaction potential (MIP) maps in [Fig. 5](#)]. Interestingly, in several cases, the negative MIP regions detected by the probe and generated by the distortion of the DNA geometry coincide with sites occupied by positively charged protein residues, LYS/ARG, at the interaction interface. Changes in structure seem to create anchoring points for cationic residues in protein tails that would

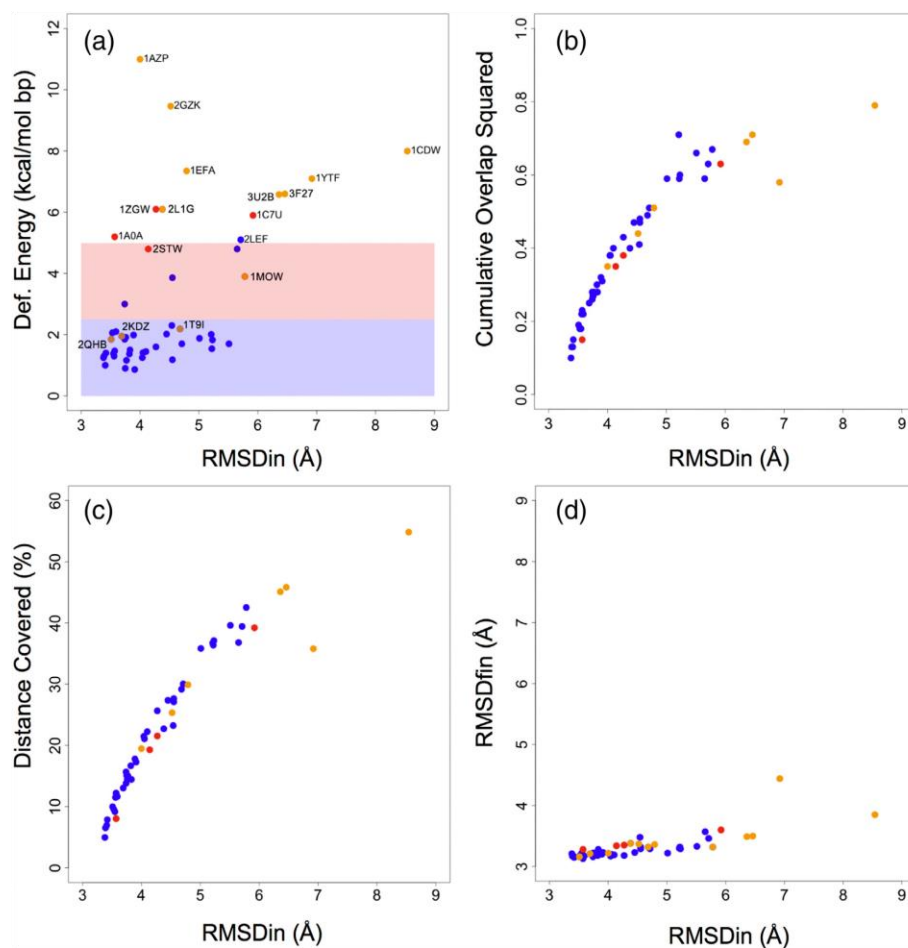


Fig. 4. Correlation between the RMSD_{in}, calculated between the average conformation along the MD simulation of the unbound DNA and experimental protein-bound structure of the DNA for the complexes studied (Table S1), and (a) deformation energy cost (kcal/mol bp) to move from the unbound to the experimental bound (bioactive) conformation in the helical space; (b) overlap squared between the essential dynamics of the unbound DNA and vector that connects the unbound and bound conformations; (c) distance covered when moving the unbound structure along the essential modes (those describing 90% of naked simulation variance) toward the bound (bioactive) structure; (d) RMSD with bound (bioactive) conformation after moving the naked structure along the essential modes in the direction of the bound (bioactive) conformation (RMSD_{fin}). The bound DNA structures detected with probable uncertainties in the experimental structure are highlighted in red, the protein DNA-complexes with DNA distorted by the protein in yellow and the remaining systems in blue dots. We marked with PDB ID names the structure with deformation energy higher than 5 kcal/mol bp.

otherwise be disordered. The analysis of the electrostatic surface of these three cases with different degrees of distortions suggests a subtle protein–DNA structural interplay where the ordered part of the protein distorts the DNA toward the bioactive state, leading to changes in the DNA electrostatic potential, which in return generates additional anchoring points for the disordered protein tails.

What is the relative prevalence of conformational selection and induced fit binding modes?

For each 1 of 174 complexes in the curated data set representative of the entire DNA–protein interactome, the mesoscopic deformation energy associated with binding was computed (see [Materials and Methods](#) and [Supplementary Methods](#)). Very interestingly, a

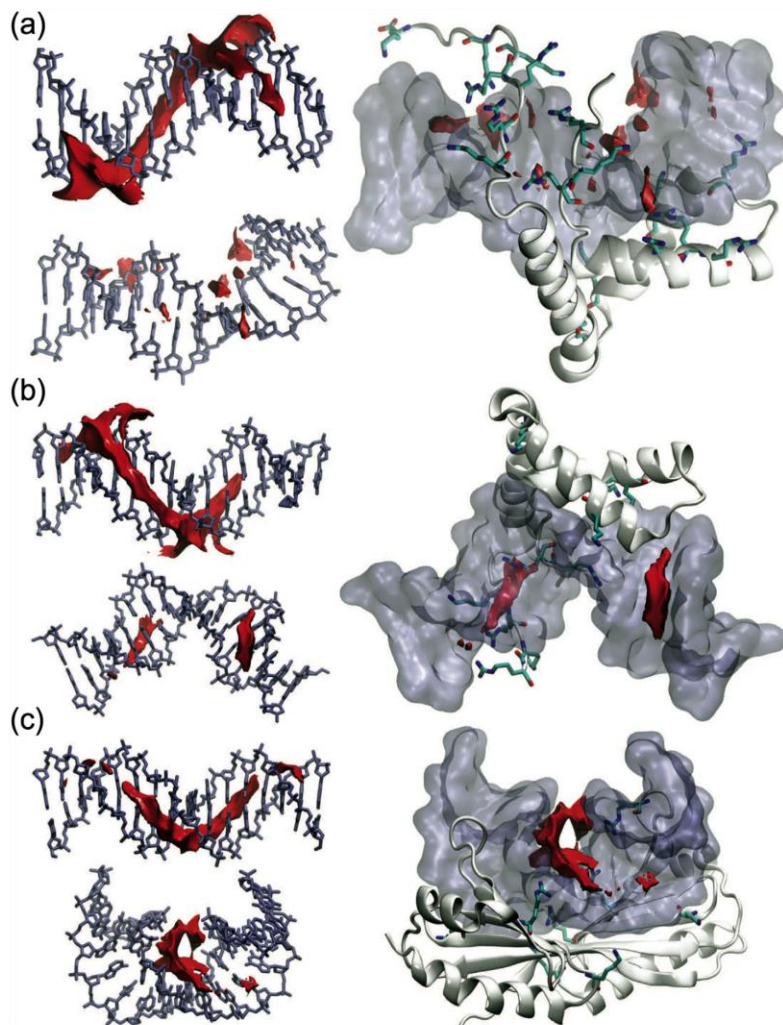


Fig. 5. MIP using Na^+ as probe for three cases, the unbound (left column, upper image) and bound (left column bottom image and right column) DNA structures. The isosurfaces (in red) have been calculated for the DNA sequences in the complexes that showed the mostly distorted structures (right column) in our data set: (a) 1J46 (isovalue = $-6.4 \text{ kcal mol}^{-1}$), (b) 3F27 (isovalue = $-7.4 \text{ kcal mol}^{-1}$), (c) 1CDW (isovalue = $-7.4 \text{ kcal mol}^{-1}$). In the right column, details of the protein residues with positive charges (lysines and arginines in licorice) pointing in the direction of the detected potential surface are represented.

vast majority of the cases follow a pure conformational selection binding mode (71%, blue area in Fig. 6), 18% of the cases fall in the twilight zone and induced fit explains around 11% of the complexes (see Fig. 6), where extremely bent or even kinked DNA is obtained by direct protein–nucleotide contacts.

Considering that the structures selected are the 17% over the entire data set of the no-redundant

PDB protein–DNA structures, this 71% corresponds to a 24% in the entire repository.

Even potential bias derived from PDB composition cannot be ruled out, our results strongly support, for complexes where the B-DNA structures are not strongly altered by the protein (mismatch/broken/unpairing, Supplementary Fig. S1), the prevalence of the conformational selection model over the induced fit

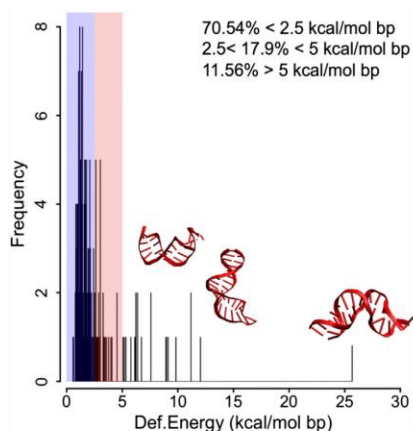


Fig. 6. Frequency of the deformation energy cost (kcal/mol bp) required moving from the unbound to the bound conformation in the helical space for all the DNA–protein interactome. In red images of structures that require high deformation energy: bent and very distorted structures at the backbone and base-pair step level (high roll value), 1YA6, 2ADW and 5H1C PDB ID respectively, from lower to higher energy. In this distribution, the number of structures that fall within the area with energy <2.5 kcal/mol bp (blue area) and energy between 2.5 and 5 kcal/mol bp (red area) are represented. Percentage for the whole selected interactome is shown in the right top corner.

one, as anticipated by atomistic simulations on the 50 selected complexes. Interestingly, these two groups are also characterized by different binding specificity. We found out that in the group defined by low energy and identified by conformational selection mechanism, a majority of the contacts are with arginine and the DNA phosphate (71%), while only 24% of the protein interacts with the bases. This protein–DNA binding is mainly driven by the electrostatics and the shape of the DNA. In the induced fit group, 34% of the interactions involve the bases, while only 51% involves the phosphates; suggesting that the protein changes the free DNA conformation to increase the interaction between the protein and the bases. This is confirmed by the increase of amidic residues present at the interface (GLN and ASN), which are very well suited to form direct bonds with the DNA bases (Supplementary Fig. S10 and scheme of recognition modes in Supplementary Fig. S11).

Furthermore, our atomistic and mesoscopic analyses suggest that in our cases DNA-interacting proteins could have evolved to recognize the native shape of the DNA duplex, avoiding the need to invest large amounts of energy in deforming the native physiological B-DNA, which would make the effector protein less efficient when competing with histones, RNAs and many other proteins.

To evaluate the generality of our conclusions, we hand-curate several (17) complexes that were excluded from the initial analysis as the PDB data set contained unpaired or modified bases (see Supplementary Methods). Supplementary Figure S12 shows that also for these complexes the energy values fall within the range expected by the conformational selection paradigm (green bars).

What is the role of base or shape readout in protein binding to cognate DNA sequences?

To answer this fundamental question, we compared the distribution of deformation energies of one million randomly generated sequences with that of the DNA sequences of our set of 50 representative complexes. Results in Fig. 7a show that the energetic cost for reaching the bioactive state for DNA sequences found in PDB is lower than for random sequences. So, it appears for these cases that the sequences in the x-ray crystal complex can reach the bioactive (bound) state much more easily than random sequences. As DNA sequences used to solve structures deposited in PDB structures tend to be consensus sequences, we can guess that, in general, the shape readout model plays a major role in selecting cognate sequences. To further validate this hypothesis, we repeated the study considering a further 20 sequences fulfilling the consensus sequence requirements taken from *in vitro* footprinting/SELEX experiments (<http://foresta.eead.csic.es/footprintdb/>) [50]. Results show again the positioning of the sequences with consensus pattern; those sequences are positioned at lower-energies with respect to the random ones (gray) and are largely favored (green lines in Fig. 7b) energetically (gray in Fig. 7b) to achieve the bound conformation. This confirms that *in vitro* high-affinity sequences are typically those showing less resistance to be distorted by the protein, as expected by the shape readout binding mechanism. In summary, theoretical results strongly favor the shape reading mechanism as a major contributor to the selection of DNA binding sequences and that nucleotide sequence alone does not fully explain the widely observed mechanism of DNA shape readout.

Materials and Methods

DNA–protein complex selection

The data set representing the DNA–protein complexes was obtained after applying a set of filters to the whole collection of protein complexes deposited in the PDB (www.rcsb.org) [51]. The initial data set was acquired from Nucleic Acid Database [52], selecting PDB entries having protein molecules attached to double-stranded B-

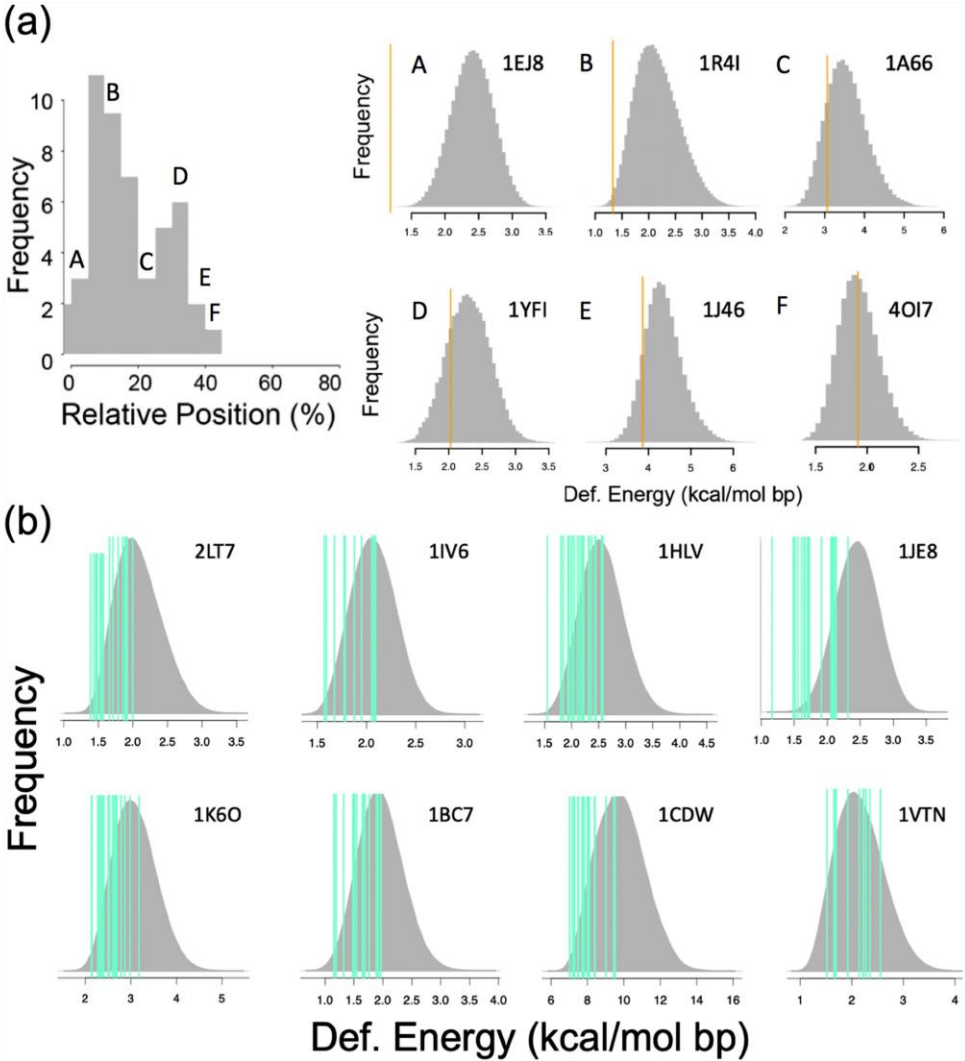


Fig. 7. (a) Relative position of the deformation energy value of the essayed PDB sequences in the frequency distribution of energies for a million random sequences (see Supplementary Methods). Highlighted with corresponding letters (A–F) are some examples, with the respective distributions on the right, identified with PDB code: in gray, the energy distribution for the random sequences and in yellow the sequence found in the experimental complex. (b) Comparison between the distribution of the deformation energies for random sequences (gray) and the deformation energy sequences with experimental high-affinity pattern (green) as found in Footprint Database.

DNA, thus avoiding single-stranded nucleic acid structures, RNA, and non-canonical B-DNA conformations. From this initial set, we removed protein redundancy and selected the 1038 unique protein–B-DNA entries found [51]. We then filtered this set excluding DNAs with modified nucleic bases, unpairing or mismatches, broken strands or non-

Watson–Crick pairing (details about PDB filtering in Supplementary Material, Supplementary Methods and Supplementary Fig. S1), obtaining a data set of 174 protein–DNA complexes that in this work defines the protein–DNA interactome. The interactions involved in the structures of this data set have been further studied using the R package

VeriNA3D [53] in Supplementary Fig. S11. From this set, we selected a sample of 50 diverse cases from the PDB (see Fig. 1 and Supplementary Table S1 for details [7]) covering different types of protein folds and function, DNA recognition modes (minor/major grooves), sequence binding motifs and structural selection. We extracted the DNA sequences from the selected 50 protein–DNA complexes, and those sequences were subjected to atomistic MD simulations in *in silico* physiological conditions.

Atomistic simulations

Starting models for all the protein-free DNAs were created using Arnott-B DNA canonical values [54]. In addition, as a way to control convergence, for a few systems trajectories were also started from the DNA conformation in the protein–DNA complex. Each system was solvated using TIP3P waters [55] in a truncated octahedron box with periodic boundary conditions, and adding Na⁺ ions until neutralization and extra salt, up to 0.15 M in NaCl, using Smith and Dang ion parameters [56]. The DNA interactions were represented using parmbsc1 force field [39–41]. All simulations were performed using Amber 14 suite of programs (AMBER 2014 San Francisco University of California). The systems were then energy minimized, thermalized and pre-equilibrated using our standard multi-step protocol [41,57] followed by 50 ns of equilibration before 0.5 μ s of unbiased MD simulations using standard simulation conditions in the NPT ensemble (see Supplementary Methods). Trajectories and associated analysis are deposited in the MuG-BigNASim database [58] (<http://mmb.irbbarcelona.org/BigNASim/>) and are freely accessible.

Analysis of trajectories

Collected trajectories (500,000 structures per system) were post-processed and analyzed using the CPPTRAJ module of the Ambertools package [59], the NAFlex server [60], VMD 1.9, Bio3D R library [61], PCAsuite [62] and Curves+ package [63], as well as “in house” software. The interaction potential (electrostatics and van der Waals) of Na⁺ and Na⁺(H₂O)₆ probes with DNA duplexes was determined using a linear approximation to the Poisson–Boltzmann equation and dielectric constant for the DNA $\epsilon_{\text{DNA}} = 8$ [64], as implemented in the CMIP program [65].

Statistical analysis of base pairs parameters

Hotelling's multivariate statistical test [66] was used to analyse whether or not the distribution of a given helical parameter in the DNA–protein com-

plex fits the expected distribution in the naked-DNA conformational ensemble. Accordingly, multivariate F statistic was defined [66] as:

$$F = \frac{n-m}{m} (\mu - \bar{x})^t S^{-1} (\mu - \bar{x}) \quad (1)$$

where μ is the vector containing the m experimental values for each base-pair step along the sequence taken from the complex structure. From the matrix ($n \times m$) containing the values for each base-pair step parameter (m) obtained from the n time frames of the MD simulations, the average values along the time (\bar{x}) and the inverse of the variance matrix (S^{-1}) have been calculated. Following Hotelling statistical test, the bound conformation is considered not sampled by the naked DNA trajectory when the computed F falls outside the confidence region $F > F_{(1-\alpha; m, n-m)}$ at $1 - \alpha = 95\%$ confidence level, where $F_{(1-\alpha; m, n-m)}$ is the quantile $1 - \alpha$ from an F distribution with $m, n - m$ degrees of freedom.

Essential dynamics analysis

Essential dynamics (ED) [67,68] analysis has been performed to determine the essential movements explaining the DNA global dynamics [67,68]. Eigenvectors (\vec{g}_i in q. 2) and eigenvalues were determined by diagonalization of the covariance matrix following the R package Bio3D [61]. The reduced set of eigenvectors that explain 90% of the variance (n in q. 2), have been selected in each case as descriptive of the essential dynamics of the duplexes. The ability of the essential dynamics of DNA to trace the conformational transition from the unbound to the bound state (given by vector \vec{R} in Eq. 2) was measured by the cumulative sum of the squared overlap (γ) between the transition vector and the eigenvectors describing the essential dynamics of the naked duplex [69,70]:

$$\gamma = \sum_{i=1}^n \left(\vec{g}_i \cdot \vec{R} \right)^2 \quad (2)$$

An additional measure of the ability of the ED space of DNA to reproduce a transition is given by the percentage of the transition (distance covered measured using the RMSD) that can be achieved by moving along the $-n$ -ED modes. In other words, how close to the bioactive conformation can the DNA arrive when moving across the easiest deformation modes (Eq. 3):

$$\text{Distance covered} = \frac{\text{RMSD}_{\text{in}} - \text{RMSD}_{90\%}}{\text{RMSD}_{\text{in}}} \% \quad (3)$$

where RMSD_{in} is calculated between the naked DNA and the protein-bound conformations. RMSD_{fin} is the minimum RMSD between the bound structure and the naked DNA after the displacement along the

3856

Sequence-Selective Protein Recognition

ED modes that describe 90% of the naked DNA motion.

Deformation energy analysis

The deformation energy associated with the DNA transition from naked to bound is approximated in the harmonic regime [71]:

$$\text{Def. Energy} = \frac{\sum_{j=1}^m E_j}{m}, \text{ with } E_j = \frac{1}{2} \sum_{s=1}^6 \sum_{t=1}^6 k_{st}^j \Delta X_s^j \Delta X_t^j \quad (4)$$

where j stands for each of the m base-pair steps of the DNA. In turn, E_j is determined from a stiffness mesoscopic model [71–73], where ΔX_s^j and ΔX_t^j are the deviation from equilibrium values in the six base-pair step helical parameters (roll, twist, tilt, slide, rise or shift) and k_{st}^j stands for the elements of the stiffness matrix obtained by inversion of the MD covariance matrix in the helical space, as determined by Olson–Lankaš model [72–74]. The equilibrium values and stiffness constants for each individual base-pair step [39,72,74] were taken from an MD simulations stored in the BigNASim [58] that cover all the unique base-pair steps in all the possible tetranucleotide environments from microsecond-long parmbsc1 simulations. Estimates of deformation energy associated with the change from the naked to the bound conformation where compared with the thermal energy fluctuation of naked B-DNA in solution. The thermal energy fluctuation of naked B-DNA is the deformation energy sampled by the DNA at room temperature. For each snapshot of each free DNA trajectory, we computed the deformation respect to the corresponding average structure. The thermal energy fluctuation is then defined taking the average plus one standard deviation from the distribution built from the collection of these energy values. The thermal energy fluctuation determined in this study from all the simulated systems amounts to 2.5 ± 0.1 kcal/mol bp.

Acknowledgments

M.O. is an ICREA (Institutió Catalana de Recerca i Estudis Avancats) academia researcher. P.D.D. is a PEDECIBA (Programa de Desarrollo de las Ciencias Básicas) and SNI (Sistema Nacional de Investigadores, Agencia Nacional de Investigación e Innovación, Uruguay) researcher. This work was supported by the Spanish Ministry of Science (Grants RTI2018-096704-B-I00), the Catalan Government (Grant 2017-SGR-134), the Instituto de

Salud Carlos III–Instituto Nacional de Bioinformática, the European Union's Horizon 2020 research and innovation program, and the Biomolecular and Bioinformatics Resources Platform (ISCIII PT 17/0009/0007 co-funded by the Fondo Europeo de Desarrollo Regional (FEDER; Grants Elixir-Excelerate: 676559 and BioExcel2: 823830 ERC:812850). Funding was also provided by the MINECO Severo Ochoa Award of Excellence from the Government of Spain (awarded to IRB Barcelona).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jmb.2019.07.021>.

Received 13 May 2019;

Received in revised form 9 July 2019;

Accepted 10 July 2019

Available online 17 July 2019

Keywords:

DNA–protein recognition;
molecular dynamics;
PDB data mining;
structural analysis;
principal component analysis

Abbreviations:

ED, essential deformation; MIP, molecular interaction potential; PDB, Protein Data Bank.

References

- [1] R. Rohs, X. Jin, S.M. West, R. Joshi, B. Honig, R.S. Mann, Origins of specificity in protein–DNA recognition, *Annu. Rev. Biochem.* 79 (2010) 233–269, <https://doi.org/10.1146/annurev-biochem-060408-091030>.
- [2] A. Tóth-Petróczy, I. Simon, M. Fuxreiter, Y. Levy, Disordered tails of homeodomains facilitate DNA recognition by providing a trade-off between folding and specific binding, *J. Am. Chem. Soc.* 131 (2009) 15084–15085, <https://doi.org/10.1021/ja9052784>.
- [3] L.-A. Harris, L.D. Williams, G.B. Koudelka, Specific minor groove solvation is a crucial determinant of DNA binding site recognition, *Nucleic Acids Res.* 42 (2014) 14053–14059, <https://doi.org/10.1093/nar/gku1259>.
- [4] A. Debnath, B. Mukherjee, K.G. Ayappa, P.K. Maiti, S.-T. Lin, Entropy and dynamics of water in hydration layers of a bilayer, *J. Chem. Phys.* 133 (2010), 174704, <https://doi.org/10.1063/1.3494115>.
- [5] G.G. Hammes, Y.-C. Chang, T.G. Oas, Conformational selection or induced fit: a flux description of reaction

- mechanism, *Proc. Natl. Acad. Sci. U. S. A.* 106 (2009) 13737–13741, <https://doi.org/10.1073/pnas.0907195106>.
- [6] H.-X. Zhou, From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions, *Biophys. J.* 98 (2010) L15–L17, <https://doi.org/10.1016/j.bpj.2009.11.029>.
- [7] J. Li, J.M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez, R. Rohs, Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding, *Nucleic Acids Res.* 45 (2017) 12877–12887, <https://doi.org/10.1093/nar/gkx1145>.
- [8] M.F. Berger, M.L. Bulyk, Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors, *Nat. Protoc.* 4 (2009) 393–411, <https://doi.org/10.1038/nprot.2008.195>.
- [9] C. Zhu, K.J.R.P. Byers, R.P. McCord, Z. Shi, M.F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M.V. Shah, M. Radhakrishnan, A.A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rofis, T. Murthy, J. Labaer, M.L. Bulyk, E. Fraenkel, R. Young, High-resolution DNA-binding specificity analysis of yeast transcription factors, *Genome Res.* 19 (2009) 556–566, <https://doi.org/10.1101/gr.090233.108>.
- [10] G. Badis, E.T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C.D. Carlson, A.J. Gossett, M.J. Hasinoff, C.L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. Li Yeo, Z.X. Yeo, N.D. Clarke, J.D. Lieb, A.Z. Ansari, C. Nislow, T.R. Hughes, A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters, *Mol. Cell* 32 (2008) 878–887, <https://doi.org/10.1016/j.molcel.2008.11.020>.
- [11] M.B. Noyes, R.G. Christensen, A. Wakabayashi, G.D. Stormo, M.H. Brodsky, S.A. Wolfe, Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites, *Cell* 133 (2008) 1277–1289, <https://doi.org/10.1016/j.cell.2008.05.023>.
- [12] A. Arvey, P. Agius, W.S. Noble, C. Leslie, Sequence and chromatin determinants of cell-type-specific transcription factor binding, *Genome Res.* 22 (2012) 1723–1734, <https://doi.org/10.1101/gr.127712.111>.
- [13] D.E. Newburger, M.L. Bulyk, UniPROBE: an online database of protein binding microarray data on protein–DNA interactions, *Nucleic Acids Res.* 37 (2009) D77–D82, <https://doi.org/10.1093/nar/gkn660>.
- [14] Z. Xie, S. Hu, S. Blackshaw, H. Zhu, J. Qian, hPDI: a database of experimental human protein–DNA interactions, *Bioinformatics* 26 (2010) 287–289, <https://doi.org/10.1093/bioinformatics/btp631>.
- [15] S. Mahony, B.F. Pugh, Protein–DNA binding in high-resolution, *Crit. Rev. Biochem. Mol. Biol.* 50 (2015) 269–283, <https://doi.org/10.3109/10409238.2015.1051505>.
- [16] T.R. Riley, M. Slattery, N. Abe, C. Rastogi, D. Liu, R.S. Mann, H.J. Bussemaker, SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes, *Methods Mol. Biol.* (2014) 255–278, https://doi.org/10.1007/978-1-4939-1242-1_16.
- [17] A. Jolma, J. Yan, T. Whittington, J. Toivonen, K.R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N.M. Luscombe, T.R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, J. Taipale, DNA-binding specificities of human transcription factors, *Cell* 152 (2013) 327–339, <https://doi.org/10.1016/j.cell.2012.12.009>.
- [18] V.A. Kuznetsov, O. Singh, P. Jenjaroenpun, Statistics of protein–DNA binding and the total number of binding sites for a transcription factor in the mammalian genome, *BMC Genomics* 11 (2010) S12, <https://doi.org/10.1186/1471-2164-11-S1-S12>.
- [19] M. Yu, G.C. Hon, K.E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, C. He, Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome, *Cell* 149 (2012) 1368–1380, <https://doi.org/10.1016/j.cell.2012.04.027>.
- [20] Z. Miao, E. Westhof, Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score, *Nucleic Acids Res.* 43 (2015) 5340–5351, <https://doi.org/10.1093/nar/gkv446>.
- [21] A. van der Vaart, Coupled binding–bending–folding: the complex conformational dynamics of protein–DNA binding studied by atomistic molecular dynamics simulations, *Biochim. Biophys. Acta - Gen. Subj.* 1850 (2015) 1091–1098, <https://doi.org/10.1016/j.bbagen.2014.08.009>.
- [22] K.M. Thayer, D.L. Beveridge, Hidden Markov models from molecular dynamics simulations on DNA, *Proc. Natl. Acad. Sci.* 99 (2002) 8642–8647, <https://doi.org/10.1073/pnas.132148699>.
- [23] Y. Zhao, D. Granas, G.D. Stormo, D. Johnson, R. Myers, Inferring binding energies from selected binding sites, *PLoS Comput. Biol.* 5 (2009), e1000590, <https://doi.org/10.1371/journal.pcbi.1000590>.
- [24] D.J. Galas, A. Schmitz, DNase footprinting: a simple method for the detection of protein–DNA binding specificity, *Nucleic Acids Res.* 5 (1978) 3157–3170 <http://www.ncbi.nlm.nih.gov/pubmed/212715> (accessed July 24, 2017).
- [25] R.B. Jones, A. Gordus, J.A. Krall, G. MacBeath, A quantitative protein interaction network for the ErbB receptors using protein microarrays, *Nature* 439 (2006) 168–174, <https://doi.org/10.1038/nature04177>.
- [26] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.-B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, R.A. Young, Transcriptional regulatory code of a eukaryotic genome, *Nature* 431 (2004) 99–104, <https://doi.org/10.1038/nature02800>.
- [27] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein–DNA interactions, *Science* (2007) 316.
- [28] R. Rohs, S.M. West, A. Sosinsky, P. Liu, R.S. Mann, B. Honig, The role of DNA shape in protein–DNA recognition, *Nature* 461 (2009) 1248–1253, <https://doi.org/10.1038/nature08473>.
- [29] G. Paillard, R. Lavery, Analyzing protein–DNA recognition mechanisms, *Structure* 12 (2004) 113–122, <https://doi.org/10.1016/j.str.2003.11.022>.
- [30] C. Chen, B.M. Pettitt, The binding process of a nonspecific enzyme with DNA, *Biophys. J.* 101 (2011) 1139–1147, <https://doi.org/10.1016/j.bpj.2011.07.016>.
- [31] A.N. Temiz, P.V. Benos, C.J. Camacho, Electrostatic hot spot on DNA-binding domains mediates phosphate desolvation and the pre-organization of specificity determinant side chains, *Nucleic Acids Res.* 38 (2010) 2134–2144, <https://doi.org/10.1093/nar/gkp1132>.
- [32] B. Bouvier, K. Zakrzewska, R. Lavery, Protein–DNA recognition triggered by a DNA conformational switch, *Angew. Chemie Int. Ed.* 50 (2011) 6516–6518, <https://doi.org/10.1002/anie.201101417>.
- [33] S. Furini, P. Barbini, C. Domene, DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence, *Nucleic*

- Acids Res. 41 (2013) 3963–3972, <https://doi.org/10.1093/nar/gkt099>.
- [34] I. Sela, D.B. Lukatsky, C. Nislow, et al., A.M. van Oijen, et al., DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity, *Biophys. J.* 101 (2011) 160–166, <https://doi.org/10.1016/j.bpj.2011.04.037>.
- [35] M. Fuxreiter, I. Simon, S. Bondos, Dynamic protein–DNA recognition: beyond what can be seen, *Trends Biochem. Sci.* 36 (2011) 415–423, <https://doi.org/10.1016/j.tibs.2011.04.006>.
- [36] L. Etheve, J. Martin, R. Lavery, Dynamics and recognition within a protein–DNA complex: a molecular dynamics study of the SKN-1/DNA interaction, *Nucleic Acids Res.* 44 (2016) 1440–1448, <https://doi.org/10.1093/nar/gkv1511>.
- [37] D.D. Boehr, R. Nussinov, P.E. Wright, The role of dynamic conformational ensembles in biomolecular recognition, *Nat. Chem. Biol.* 5 (2009) 789–796, <https://doi.org/10.1038/nchembio.232>.
- [38] T.-P. Chiu, F. Comoglio, T. Zhou, L. Yang, R. Paro, R. Rohs, DNASHapeR: an R/Bioconductor package for DNA shape prediction and feature encoding, (n.d.), doi:<https://doi.org/10.1093/bioinformatics/btv735>.
- [39] I. Ivani, P.D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J.L. Gelpi, C. González, M. Vendruscolo, C.A. Loughton, S.A. Harris, D.A. Case, M. Orozco, Parmbsc1: a refined force field for DNA simulations, *Nat. Methods* 13 (2015) 55–58, <https://doi.org/10.1038/nmeth.3658>.
- [40] P.D. Dans, I. Ivani, A. Hospital, G. Portella, C. González, M. Orozco, How accurate are accurate force-fields for B-DNA? *Nucleic Acids Res.* 45 (2017), gkw1355. <https://doi.org/10.1093/nar/gkw1355>.
- [41] P.D. Dans, L. Danilăne, I. Ivani, T. Dršata, F. Lankaš, A. Hospital, J. Walther, R.I. Pujagut, F. Battistini, J.L. Gelpi, R. Lavery, M. Orozco, Long-timescale dynamics of the Drew–Dickerson dodecamer, *Nucleic Acids Res.* 44 (2016) 4052–4066, <https://doi.org/10.1093/nar/gkw264>.
- [42] A. Kuzmanic, P.D. Dans, M. Orozco, An in-depth look at DNA crystals through the prism of molecular dynamics simulations, *Chem.* (2019) <https://doi.org/10.1016/j.CHEMPR.2018.12.007>.
- [43] P.D. Dans, J. Walther, H. Gómez, Multiscale simulation of DNA, *Curr. Opin. Struct. Biol.* 37 (2016) 29–45, <https://doi.org/10.1016/j.SBI.2015.11.011>.
- [44] L.J.W. Murray, W.B. Arendall, D.C. Richardson, J.S. Richardson, RNA backbone is rotameric, *Proc. Natl. Acad. Sci. U. S. A.* 100 (2003) 13904–13909, <https://doi.org/10.1073/pnas.1835769100>.
- [45] A. Ben Imeddourene, X. Xu, L. Zargarian, C. Oguey, N. Follpe, O. Mauffret, B. Hartmann, The intrinsic mechanics of B-DNA in solution characterized by NMR, *Nucleic Acids Res.* 44 (2016) 3432–3447, <https://doi.org/10.1093/nar/gkw084>.
- [46] R. Galindo-Murillo, J.C. Robertson, M. Zgarbová, J. Šponer, M. Otyepka, P. Jurečka, T.E. Cheatham, Assessing the current state of Amber force field modifications for DNA, *J. Chem. Theory Comput.* 12 (2016) 4114–4127, <https://doi.org/10.1021/acs.jctc.6b00186>.
- [47] P.D. Dans, A. Pérez, I. Faustino, R. Lavery, M. Orozco, Exploring polymorphisms in B-DNA helical conformations, *Nucleic Acids Res.* 40 (2012) 10668–10678, <https://doi.org/10.1093/nar/gks884>.
- [48] P.D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery, M. Orozco, Unraveling the sequence-dependent polymorphic behavior of d (CpG) steps in B-DNA, *Nucleic Acids Res.* 42 (2015) 11304–11320.
- [49] N. Abe, I. Dror, L. Yang, M. Slattery, T. Zhou, H.J. Bussemaker, R. Rohs, R.S. Mann, Deconvolving the recognition of DNA shape from sequence, *Cell.* 161 (2015) 307–318, <https://doi.org/10.1016/j.CELL.2015.02.008>.
- [50] B. Contreras-Moreira, 3D-footprint: a database for the structural analysis of protein–DNA complexes, *Nucleic Acids Res.* 38 (2010) D91–D97, <https://doi.org/10.1093/nar/gkp781>.
- [51] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, *Nucleic Acids Res.* 28 (2000) 235–242, <https://doi.org/10.1093/nar/28.1.235>.
- [52] B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A.I. Petrov, B. Sweeney, C.L. Zirbel, N.B. Leontis, H.M. Berman, The nucleic acid database: new features and capabilities, *Nucleic Acids Res.* 42 (2014) D114–D122, <https://doi.org/10.1093/nar/gkt980>.
- [53] D. Gallego, L. Darré, P.D. Dans, M. Orozco, VeriNA3d: an R package for nucleic acids data mining, *Bioinformatics.* (2019) <https://doi.org/10.1093/bioinformatics/btz553>.
- [54] S. Arnott, D.W. Hukins, Optimised parameters for A-DNA and B-DNA, *Biochem. Biophys. Res. Commun.* 47 (1972) 1504–1509 <http://www.ncbi.nlm.nih.gov/pubmed/5040245> (accessed May 22, 2017).
- [55] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.* 79 (1983) 926–935, <https://doi.org/10.1063/1.445869>.
- [56] D.E. Smith, L.X. Dang, Computer simulations of NaCl association in polarizable water, *J. Chem. Phys.* 100 (1994) 3757–3766, <https://doi.org/10.1063/1.466363>.
- [57] A. Pérez, F.J. Luque, M. Orozco, Dynamics of B-DNA on the microsecond time scale, *J. Am. Chem. Soc.* 129 (2007) 14739–14745, <https://doi.org/10.1021/ja0753546>.
- [58] A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra, P. D. Dans, F. Battistini, J. Torres, R. Goñi, M. Orozco, J.L. Gelpi, BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data, *Nucleic Acids Res.* 44 (2016) D272–D278, <https://doi.org/10.1093/nar/gkv1301>.
- [59] D.R. Roe, T.E. Cheatham, PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data, *J. Chem. Theory Comput.* 9 (2013) 3084–3095, <https://doi.org/10.1021/ct400341p>.
- [60] A. Hospital, I. Faustino, R. Collepardo-Guevara, C. Gonzalez, J. L. Gelpi, M. Orozco, NAFlex: a web server for the study of nucleic acid flexibility, *Nucleic Acids Res.* 41 (2013) W47–W55, <https://doi.org/10.1093/nar/gkt378>.
- [61] L. Skjærven, X.-Q. Yao, G. Scarabelli, B.J. Grant, Integrating protein structural dynamics and evolutionary analysis with Bio3D, *BMC Bioinformatics.* 15 (2014) 399, <https://doi.org/10.1186/s12859-014-0399-6>.
- [62] T. Meyer, C. Ferrer-Costa, A. Pérez, M. Rueda, A. Bidon-Chanal, F.J. Luque, A. Charles, A. Loughton, M. Orozco, Essential dynamics: a tool for efficient trajectory compression and management, (2006). doi:<https://doi.org/10.1021/CT050285B>.
- [63] R. Lavery, M. Moakher, J.H. Maddocks, D. Petkeviciute, K. Zakrzewska, Conformational analysis of nucleic acids revisited: Curves+, *Nucleic Acids Res.* 37 (2009) 5917–5929, <https://doi.org/10.1093/nar/gkp608>.
- [64] A. Cuervo, P.D. Dans, J.L. Carrascosa, M. Orozco, G. Gomila, L. Fumagalli, Direct measurement of the dielectric polarization properties of DNA, *Proc. Natl. Acad. Sci.* 111

- (2014) E3624–E3630, <https://doi.org/10.1073/pnas.1405702111>.
- [65] J.L. Gelpí, S.G. Kalko, X. Barril, J. Cirera, X. de La Cruz, F.J. Luque, M. Orozco, Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins, *Proteins*. 45 (2001) 428–437 <http://www.ncbi.nlm.nih.gov/pubmed/11746690> (accessed May 24, 2017).
- [66] W. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*, n.d.
- [67] M. Orozco, A. Pérez, A. Noy, F.J. Luque, Theoretical methods for the simulation of nucleic acids, *Chem. Soc. Rev.* 32 (2003) 350–364, <https://doi.org/10.1039/B207226M>.
- [68] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Essential dynamics of proteins, *Proteins Struct. Funct. Genet.* 17 (1993) 412–425, <https://doi.org/10.1002/prot.340170408>.
- [69] A. Pérez, J.R. Blas, M. Rueda, J.M. López-Bes, X. de la Cruz, M. Orozco, Exploring the essential dynamics of B-DNA, *J. Chem. Theory Comput.* 1 (2005) 790–800, <https://doi.org/10.1021/ct050051s>.
- [70] A. Noy, F. Javier Luque, M. Orozco, Theoretical analysis of antisense duplexes: determinants of the RNase H susceptibility, (2008). doi:<https://doi.org/10.1021/JA076734U>.
- [71] G. Portella, F. Battistini, M. Orozco, Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations, *PLoS Comput. Biol.* 9 (2013), e1003354. <https://doi.org/10.1371/journal.pcbi.1003354>.
- [72] F. Lankas, J. Sponer, J. Langowski, T.E. Cheatham, DNA basepair step deformability inferred from molecular dynamics simulations, *Biophys. J.* 85 (2003) 2872–2883, [https://doi.org/10.1016/S0006-3495\(03\)74710-9](https://doi.org/10.1016/S0006-3495(03)74710-9).
- [73] A. Pérez, F. Lankas, F.J. Luque, M. Orozco, Towards a molecular dynamics consensus view of B-DNA flexibility, *Nucleic Acids Res.* 36 (2008) 2379–2394, <https://doi.org/10.1093/nar/gkn082>.
- [74] W.K. Olson, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, *Proc. Natl. Acad. Sci.* 95 (1998) 11163–11168, <https://doi.org/10.1073/pnas.95.19.11163>.

Bibliography for Chapter 3

- [1] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson, "Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids," *Nature*, vol. 171, no. 4356, pp. 738–740, Apr. 1953.
- [2] R. E. Franklin and R. G. Gosling, "Molecular Configuration in Sodium Thymonucleate," *Nature*, vol. 171, no. 4356, pp. 740–741, Apr. 1953.
- [3] T. Maehigashi, C. Hsiao, K. Kruger Woods, T. Moulaei, N. V. Hud, and L. Dean Williams, "B-DNA structure is intrinsically polymorphic: even at the level of base pair positions," *Nucleic Acids Res.*, vol. 40, no. 8, pp. 3714–3722, Apr. 2012.
- [4] J. Kypr, I. Kejnovska, D. Renciuik, and M. Vorlickova, "Circular dichroism and conformational polymorphism of DNA," *Nucleic Acids Res.*, vol. 37, no. 6, pp. 1713–1725, Jan. 2009.
- [5] P. D. Dans, I. Faustino, F. Battistini, K. Zakrzewska, R. Lavery, and M. Orozco, "Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA," *Nucleic Acids Res.*, vol. 42, no. 18, pp. 11304–11320, Oct. 2014.
- [6] A. B. Imeddourene *et al.*, "The intrinsic mechanics of B-DNA in solution characterized by NMR," *Nucleic Acids Res.*, vol. 44, no. 7, pp. 3432–3447, Apr. 2016.
- [7] Y. Tian, M. Kayatta, K. Shultis, A. Gonzalez, L. J. Mueller, and M. E. Hatcher, "³¹P NMR Investigation of Backbone Dynamics in DNA Binding Sites [†]," *J. Phys. Chem. B*, vol. 113, no. 9, pp. 2596–2603, Mar. 2009.
- [8] P. D. Dans, A. Pérez, I. Faustino, R. Lavery, and M. Orozco, "Exploring polymorphisms in B-DNA helical conformations," *Nucleic Acids Res.*, vol. 40, no. 21, pp. 10668–10678, Nov. 2012.
- [9] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein-DNA crystal complexes," *Proc. Natl. Acad. Sci.*, vol. 95, no. 19, pp. 11163–11168, Sep. 1998.
- [10] F. Lankaš, J. Šponer, J. Langowski, and T. E. Cheatham, "DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations," *Biophys. J.*, vol. 85, no. 5, pp. 2872–2883, Nov. 2003.
- [11] R. Lavery *et al.*, "A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA," *Nucleic Acids Res.*, vol. 38, no. 1, pp. 299–313, Jan. 2010.
- [12] I. Ivani *et al.*, "Parmbsc1: a refined force field for DNA simulations," *Nat. Methods*, vol. 13, no. 1, pp. 55–58, Jan. 2016.
- [13] S. B. Dixit *et al.*, "Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. II: Sequence Context Effects on the Dynamical Structures of the 10 Unique Dinucleotide Steps," *Biophys. J.*, vol. 89, no. 6, pp. 3721–3740, Dec. 2005.
- [14] M. Pasi *et al.*, "μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA," *Nucleic Acids Res.*, vol. 42, no. 19, pp. 12272–12283, Oct. 2014.
- [15] A. Hospital *et al.*, "BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D272–D278, Jan. 2016.
- [16] P. D. Dans *et al.*, "The physical properties of B-DNA beyond Calladine-Dickerson rules."
- [17] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000.

Chapter 4 . Determinants of nucleosome architecture in yeast

In the previous chapter we described how the protein-DNA binding is highly dependent on the physical properties and intrinsic flexibility of the DNA. In the following we will focus on a very important protein-DNA complex, the nucleosome, that is present genome-wide and crucial for the regulation of gene expression in the cell. We study the role of DNA physical properties to define clear regions of nucleosome depletion as well as other determinants of nucleosome positioning in *Saccharomyces cerevisiae*.

As previously reported ([1], [2]), the correlation between gene expression and nucleosome architecture at promoters is high. Typically, there is a clear nucleosome free region (NFR) around the transcription start site (TSS) of actively transcribed genes, having a strongly positioned +1 nucleosome and becoming fuzzier as nucleosomes are further downstream the TSS [3]–[6]. On the other side, at the transcription termination sites (TTS) the existence of these nucleosome depletion signals is controversial. Our group and others supported the existence of NFRs around TTS [5], [7], linked to the unusual physical properties at these loci, while others claimed that NFRs at TTS are an artefact of the small distance between the TSS of neighboring genes[8]. Here, we evaluated the nucleosome organization at the 3' end of genes from MNase-seq data, but centering at the -last nucleosome (last nucleosome within the gene body, upstream from the TTS) instead of the TTS (see

Figure 4.1), finding that an NFR downstream of the -last nucleosome is present both when there is a nearby TSS (tandem oriented genes) or TTS (convergent orientation). Then we conclude that there is an area depleted of nucleosomes at the 3' end of the genes.

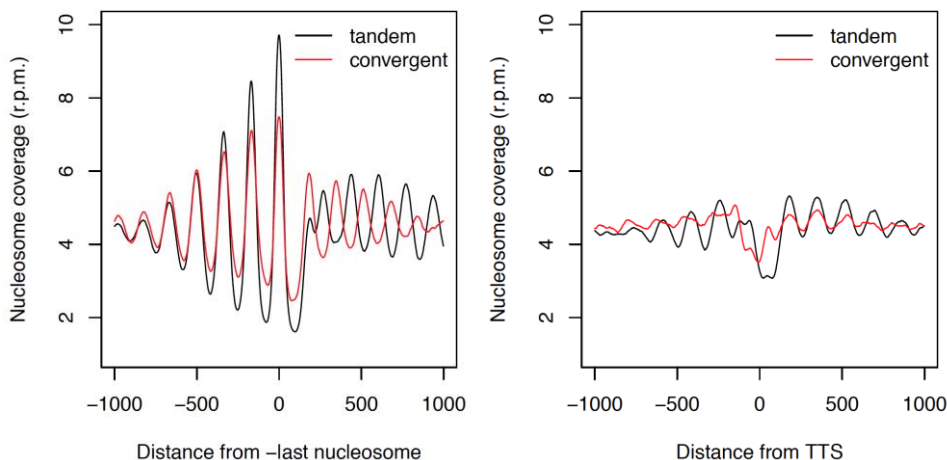


Figure 4.1. Nucleosome coverage, averaged among all genes, centered around the -last nucleosome (left panel) or the transcription termination site (TTS). Separate curves are shown according to the orientation of the downstream gene: tandem (black line, $\rightarrow\rightarrow$) or convergent (red line, $\rightarrow\leftarrow$).

Next, we explored the possibility to predict nucleosome architecture along the gene body by statistical positioning, leading to the maximum occupancy of nucleosomes in between two well-positioned nucleosomes in the vicinities of the TSS (the +1 nucleosome) and the TTS (the -last nucleosome). Particularly, we tested the ability of a simple signal transduction model with two emitters (at the +1 and -last positions) and a periodic distance-decay signal. We demonstrated that this simple model, with a periodicity of 165 bp (for yeast), can predict with high accuracy the nucleosome coverage at gene bodies (see an example for one gene in **Figure 4.2**). The high predictive power of this simple model supports the idea that once the strong signals (intrinsic or protein-mediated) are responsible for positioning the +1 and -last nucleosomes, simple statistical positioning explains the nucleosome distribution along the gene body.

From the experimental and predicted intra-genic nucleosome coverages, we observed two classes of genes according to their nucleosome coverage profile: a set of genes where the two signals from the +1 and -last nucleosomes overlap significantly and the nucleosomes tend to be well-positioned (phased genes), and a second set of genes where the two signals are not in phase and the nucleosomes along the gene body are fuzzier (unphased genes).

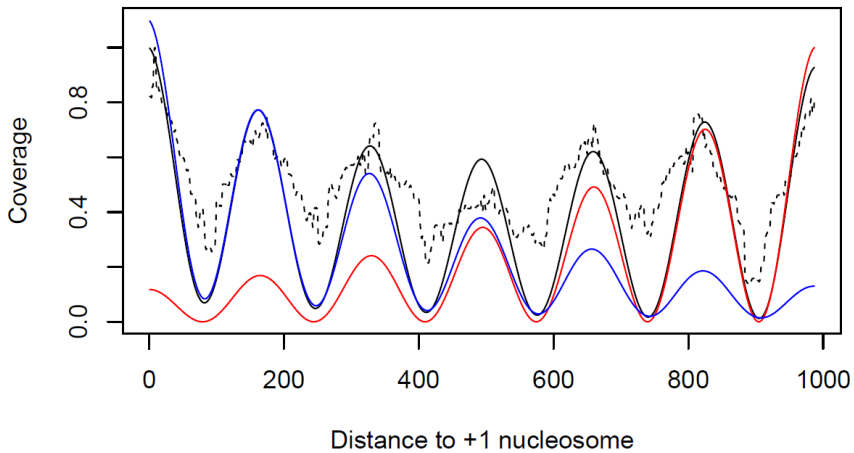


Figure 4.2. Example of prediction of nucleosome coverage for a gene body. Experimental normalized nucleosome coverage (black dotted line) for YOR039W gene between the +1 and the -last nucleosomes. The predicted coverage (black continuous line) is computed by combining the signal emitted from the +1 nucleosome (blue line) and from the -last nucleosome (red line).

We experimentally explored the effect of phasing on the gene body by adding an 81-nucleotide (81-nt) sequence to eight selected genes: four phased genes and four control not phased genes. The experiment revealed that the nucleosome organization changed as consequence of the sequence addition, obtaining fuzzier and less periodic nucleosomes in the originally phased group, but negligible changes in the control genes.

We found that genes with well-located nucleosomes in the gene body tend to have a larger expression level than those with fuzzy nucleosomes, which would suggest a causal relationship: ordered nucleosomes in the gene body leads to higher expression. To check this hypothesis, we analyzed the impact of the addition of the 81-nt sequence, finding little effect on transcription levels of the eight genes,

including those originally phased. The lack of effect of periodicity on gene expression led us to examine the opposite relation: is nucleosome periodicity affected by transcription? For this, we performed MNase-seq experiments in cells treated by 1,10-phenanthroline, a metal chelator that stalls the polymerase at the promoters and stops transcription [9]–[11]. We observed that addition of 1,10-phenanthroline leads to larger NFRs (mostly from -1 nucleosome displacement, see **Figure 4.3**), an increase in the proportion of fuzzy nucleosomes and a decrease in the proportion of phased genes. This strongly suggests that it is the presence of RNA polymerase that affects nucleosome architecture and not the reverse.

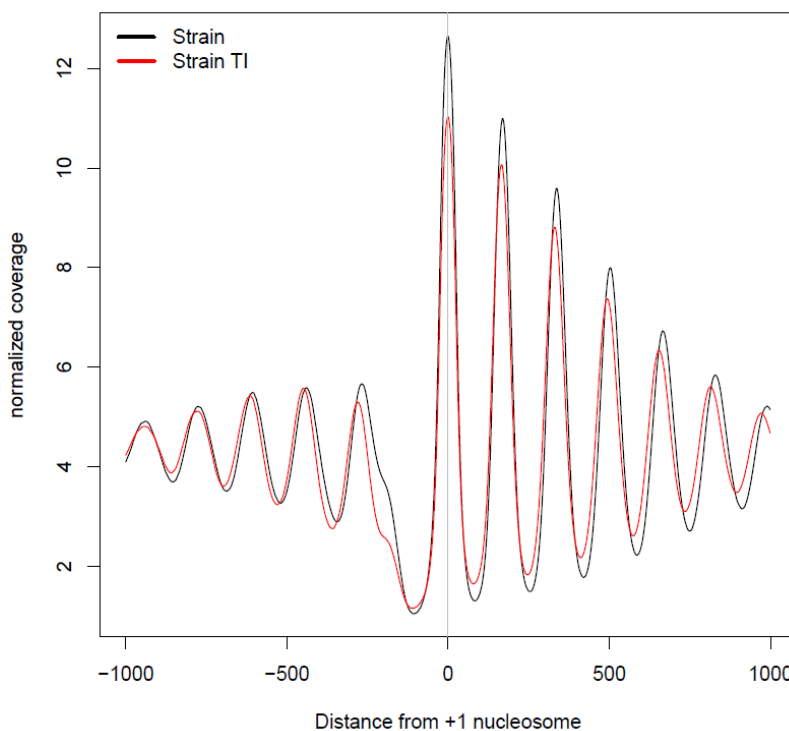


Figure 4.3. Effect of transcription inhibition on nucleosome coverage. Nucleosome coverage, normalized to reads per million, in a strain treated by 1,10 phenanthroline (red curve) and in the control strain (black line). The average coverage among all genes, centered at +1 nucleosome is shown.

Finally, since we observed that the two NFRs at the 5' and 3' ends of the genes are important to define the nucleosome architecture, we investigated the determinants

of the observed nucleosome depletion at those loci. We hypothesized that nucleosome architecture is a combination of intrinsic (sequence-dependent) and extrinsic (DNA-binding proteins, transcriptional or replication machinery, nucleosome remodelers) factors in the nucleus. We used the deformation energy of the DNA to form a nucleosome derived from physical descriptors to model intrinsic effects, and the predicted transcription factor binding site (TFBS) affinity to model extrinsic effects (see Chapter 2 for their definition). With these two variables, we built a machine learning classifier for NFR prediction along the yeast genome. We trained an ExtraTrees [12] predictor using all the TSS-NFRs and TTS-NFRs defined by well-positioned nucleosomes, except those regions in *chrI*, which were used for testing the performance of the classifier, obtaining an Area Under the Curve (AUC) of 0.77 in the test set, which include an entire chromosome. Hence, the position of many NFRs in the yeast genome can be explained by the high stiffness of the DNA sequence and the presence of binding proteins that compete with nucleosomes.

Publication:

Diana Buitrago^{*}, Mireia Labrador^{*}, Pau De Jorge, Federica Battistini, Isabelle Brun Heath and Modesto Orozco. The interplay between periodicity, DNA physical properties and effector binding define nucleosome architecture in yeast (*in preparation*).

Supplementary material for this article can be found in the **Annex III**.

^{*} Equally contributing authors

THE INTERPLAY BETWEEN PERIODICITY, DNA PHYSICAL PROPERTIES AND EFFECTOR BINDING DEFINE NUCLEOSOME ARCHITECTURE IN YEAST

Diana Buitrago^{1,&}, Mireia Labrador^{1,&}, Pau De Jorge¹, Federica Battistini¹, Isabelle Brun Heath¹ and Modesto Orozco^{1,2*}

We explored the role of periodicity, DNA physical properties and the binding of effector proteins in the positioning of nucleosomes in *Saccharomyces Cerevisiae*. We found that the positions of the first and the last nucleosomes in a gene are the result of a combination of physical properties and binding of effector proteins that define clear eviction signals at the extremes of the genes. On the contrary, nucleosomes in the gene body are placed by distance-decaying periodic signals emitted by the +1 and -last nucleosomes in the gene. Wide nucleosome free regions and periodic nucleosome strings are characteristic of active genes, proving a correlation between nucleosome architecture and gene structure. A variety of experiments demonstrate transcriptional activity is more a reason for than a consequence of nucleosome architecture.

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain; ²Departament de Bioquímica i Biomedicina, Universitat de Barcelona, Barcelona, Spain

& These authors contributed equally to this work

* Correspondence to M.Orozco: modesto.orozco@irbbarcelona.org

INTRODUCTION

Nucleosome (the basic unit of eukaryotic chromatin) is formed by 147bp of DNA wrapped around an octamer of histones (Richmond and Davey 2003), followed by a linker where, in complex eukaryotic organisms, an additional histone (H1) is bound (Izzo et al. 2008). Nucleosomes are not randomly placed in the genome, but maintain well defined and conserved positioning (Yuan et al. 2005; Mavrich et al. 2008b; Shivaswamy et al. 2008; Valouev et al. 2011), marked by the presence of nucleosome free regions (NFRs) which determine the boundaries of strings of nucleosomes (Mavrich et al. 2008a; Vaillant et al. 2010). Most significant NFRs (are associated with the promoters of genes (upstream the Transcription Start Sites; TSSs), the replication origins (ORIs) and the Transcription Termination Sites; TTSS; see Deniz et al., 2011, 2016). General consensus is that at promoters, NFRs are regions preferentially recognized by effector proteins involved in the regulation of gene activity and, accordingly, wide and well defined NFRs are typically associated to active chromatin (Weiner et al. 2010).

Over the last two decades, many efforts have been made to discover the main determinants of nucleosome positioning (reviewed in (Jiang and Pugh 2009; Clark 2010; Struhl and Segal 2013; Lieleg et al. 2015; Chereji and Clark 2018). Some authors have suggested that DNA physical properties are the most important factor to define nucleosome positioning, with NFRs being located at regions where the mechanical cost of wrapping DNA around nucleosomes is very high (Suter et al. 2000; Kaplan et al. 2009; Deniz et al. 2011). On the contrary, others have suggested that nucleosome positioning is dictated by cellular machinery involving a complex interplay of chromatin remodelers, transcription factors and RNA Polymerase activity (Hughes et al. 2012; Lorch et al. 2014; Kubik et al. 2019). Accurate chromatin reconstitution experiments (Zhang and Pugh 2011; Krietenstein et al. 2016) demonstrated that NFRs are well reproduced in *in vitro* reconstitution experiments performed in the absence of any cellular machinery, but the exact boundaries of the NFRs could only be achieved if a cell-free extract is added. These findings suggest that while physical principles can signal NFRs, cellular machinery is required for the correct definition of the NFR boundaries (Zhang and Pugh 2011; Struhl and Segal 2013; Kubik et al. 2019). The same conclusions on the dual role of extrinsic and intrinsic factors in nucleosome positioning can be reached by analyzing perturbation in nucleosome architectures associated to stress, changes in cell cycle, the source of nutrients, or the cell metabolic cycle (Shivaswamy et al. 2008; Kaplan et al. 2009; Deniz et al. 2016; Nocetti and Whitehouse 2016).

We introduce here an additional element in the debate between intrinsic and extrinsic determinants for nucleosome positioning: the role of periodicity. We demonstrate that by using signal transmitter theory with two emitters centered at the first (+1) and the last (-last) nucleosomes, located adjacent to the NFRs at the beginning and the end of the genes, the nucleosome architecture can be very well reproduced. We also found that positions of NFRs can be captured by machine learning (ML) methods using sequence-dependent descriptors of DNA physical properties and TRANSFAC (Wingender et al. 1996) annotations of transcription

factor binding sites (TFBS). We found a clear correlation between nucleosome architecture and gene activity with wide NFRs and periodic nucleosome arrays signaling active genes. However, synthetic experiments demonstrate that changes in phasing of nucleosomes leading to alterations in the periodicity of nucleosomes in the gene body do not affect gene activity. We also prove that in conditions of inhibition of RNA polymerase progression NFRs at TSS become wider and phasing of nucleosomes in the gene body is decreased. The causality in the correlation between nucleosome architecture and gene expression is then deciphered: gene activity influences more nucleosome architecture than in the reverse.

METHODS

Yeast strains and growth conditions. *Saccharomyces cerevisiae* PPY1 strain (*MATa his3 Δ 0 leu2 Δ 0 met15 Δ 0 ura3 Δ 0 bar1::leu2*) was transformed with the appropriate DNA fragments to generate all the mutant strains used in this work. PPY1 strain was obtained from Oscar Aparicio's lab at the University of Southern California, USA. For the selection of the mutant strains, we used YPD with or without 5-FOA (5-Fluoroorotic acid) and SD (Synthetic Defined) with the required amino acids.

Mutant strains generation. We generated 4 mutant strains, with the 81-nt DNA sequence (5'-GCGTGTGTGTTTTCTCCGAGGAGAAACATTCAAATCTTGTGCTATGGCTTTCCTACCGTCTG CGCCATCCATCTTTTCGC-3') inserted in the coding sequence of 2 selected genes/strain (see Suppl. Table S1). Target genes were selected as they are not essential and show phased nucleosomes (UBX5, CKB2, PPT1, TRP4; see definition below) and four control genes which were not-phased (BSP1, DGK1, SLM3, PAN5) (see Suppl. Methods). The insert was placed in linkers to avoid direct interference to specific nucleosome (Suppl. Table S1). The 81-nt sequence was selected, as it did not match any existing yeast sequence, and did not favor nor disfavor nucleosome formation (see Suppl. Figure S1), or affect reading frame. The strains were produced using the *Delitto perfetto* strategy described in (Storici and Resnick 2006).

RNA extraction and RT-qPCR. Exponential cultures were arrested at late G1 by alpha-factor. RNA was obtained from 20ml yeast cultures (OD₆₀₀ 0.8) using the hot-phenol method (see Suppl. Methods). cDNA synthesis was done with the First Strand cDNA Synthesis Kit (Roche) using oligo dT and following the provider instructions. Gene expression levels were determined by quantitative PCR using the LightCycler 480 sybr green I master (Roche). The oligonucleotides used for the qPCR are listed in Suppl. Table S2.

Transcription inhibition. In order to determine the correct incubation time to inhibit transcription without killing the cells, we selected 2 genes with low RNA stability (RPA135 and NMD3) and 2 genes with high RNA stability (ACT1 and DGK1) to serve as controls (Grigull et al. 2004). We then measured their mRNA level by qPCR after incubation with 10-phenanthroline at 100 μ g/ml at 30°C during 0, 5, 15, 30 and 45 minutes. Using this approach, we observed that the amount of RPA135 and NMD3 mRNA started to decrease after 30 min.

This incubation time was selected to perform MNase-seq experiment on cells with inhibited transcription.

MNase digestion. The Micrococcal nuclease (MNase) digestion was performed on semi-intact yeast cells prepared as described (Schlenstedt et al. 1993). We optimised the MNase digestion conditions for each sample to obtain about 80% of mononucleosomes (Suppl. Methods). The integrity and size distribution of digested fragments were determined using the microfluidics-based platform Bioanalyzer (Agilent) prior to sample preparations and sequencing. The sample preparation was done using the Illumina TruSeq DNA sample preparation kit for whole genome sequencing, following Illumina standard protocol. The libraries were sequenced paired-end on a HiSeq2000, v4, 2x75bp, with approximately 10 M PE reads/sample.

Nucleosome calling. MNase-seq paired-end reads were mapped to customized versions of yeast genome (SacCer3, UCSC), containing the inserted sequences in the modified genes, using Bowtie (Langmead et al. 2009) aligner, allowing up to two mismatches. Output files were imported in R, reads were trimmed to 50bp maintaining the original center and transformed to reads per million. Peak calling was performed, after noise filtering, with nucleR package implemented in the Nucleosome Dynamics platform (Flores and Orozco 2011; Buitrago et al., 2019) using the parameters: peak width of 147 bp, peak detection threshold of 35%, maximum overlap of 80 bp. Nucleosome calls were considered well-positioned when nucleR peak width score and height score were higher than 0.6 and 0.4, respectively, and fuzzy otherwise.

Nucleosome periodicity and phasing. Periodicity in the nucleosome positioning was determined for each gene by computing the autocorrelation coefficient, as proposed in (Wan et al. 2009); see eq. 1:

$$R(T) = \int_{X_1}^{X_2} I(x) \cdot I(x - T) dx \quad (1)$$

where X_1 and X_2 stand for the limits of a sampling window (e.g. the position of TSS and TTS), I is the function representing nucleosome coverage for all genes and T is a putative periodicity. Autocorrelation coefficients for different periods were normalized as shown in eq. 2:

$$\hat{R}(T) = \frac{R(T)}{R(0)} \quad (2)$$

Nucleosome period is defined as the value of T that optimizes $\hat{R}(T)$. Periodic genes are those showing large autocorrelation coefficient values (eq. 1). Phased genes are defined as those where the +1 to -last distance (L) is a multiple of the period (T). Unphased genes are those

where the distance from integer (DFI) score (eq. 3), defined as the modulus of the ratio length/periodicity, is close to $T/2$. Not-phased genes refer to intermediate values.

$$DFI = L - T \cdot \text{round}\left(\frac{L}{T}\right) \quad (3)$$

Signal transmission theory for nucleosome positions. We propose a simple signal decay model, where the coverage at a given position is given by the addition of two positioning signals, one starting from the +1 and another from the -last nucleosome (see eq. 4-5 below):

$$Cov^{+1}(X) = \left(1 + \alpha + \sin\left(\frac{\pi}{2} + 2 \cdot \frac{\pi}{T} X\right)\right) \sigma^{\left(\frac{|X|}{T}\right)} \quad (4)$$

$$Cov^{-last}(X') = \left(1 + \sin\left(\frac{\pi}{2} + 2 \cdot \frac{\pi}{T} X'\right)\right) \sigma^{\left(\frac{|X'|}{T}\right)} \quad (5)$$

where X is the distance from +1 nucleosome and X' is the distance from -last nucleosome, $X'=L-X$. The shifting factor α corrects for the higher density of reads at the +1 nucleosome and the decay factor σ accounts for the reduced coverage as we move away from the NFR. We evaluated different values and selected those that maximized the correlation between the observed experimental coverage and the predicted (α was set to 0.2 and σ was set to 0.7). The total coverage is then normalized to guarantee an effective decay of the signal:

$$Cov(X) = \frac{Cov^{+1}(X) + Cov^{-last}(X')}{Cov^{+1}(0) + Cov^{-last}(0)} \quad (6)$$

where, $Cov^{+1}(0)$, $Cov^{-last}(0)$ are the values of the two emitting signals at the +1 nucleosome dyad, which are used as denominator to normalize the $Cov(X)$ to 1 at this position.

Deformation energy, elastic energy associated to the DNA deformation from the naked to the nucleosomal DNA was calculated in the harmonic regime using (eq. 7):

$$Def. Energy = \frac{\sum_{j=1}^{147} E_j}{147}, \text{ with } E_j = \frac{1}{2} \sum_{s=1}^6 \sum_{t=1}^6 k_{st}^j \Delta X_s^j \Delta X_t^j \quad (7)$$

where j stands for each of the 147 base pair steps of the DNA stretches. E_j is the elastic energy required at each base pair step determined using the stiffness matrix (K), and ΔX_s^j and ΔX_t^j are the differences between the nucleosome and equilibrium values for the 6 base pair step helical parameters (roll, twist, tilt, slide, rise or shift). The equilibrium values and stiffness constants for each individual base pair step were taken from MD simulations that cover all the unique base pair steps in all the possible tetranucleotide environments from microsecond-long parmbc1 simulations (Dans et al. In press; Walther et al. In press).

NFR prediction with randomized decision trees. We trained Random Forest (Breiman 2001) and Extremely randomized trees (Extra-Trees, (Geurts et al. 2006) classifiers to predict NFRs using as predictive features the deformation energy and TFBS density profiles around a given position. Data on nucleosome distributions were obtained from MNase-seq experiments of Yeast synchronized at the G1 phase and processed with nucleR (Flores and Orozco 2011) as implemented in Nucleosome Dynamics (Buitrago et al. 2019). Data on Chromosomes II-XVI were used for training, while chromosome I was used to test the final model. The training set was divided into two portions: the learning subset (75% of data) and the validation subset (25%), which was used to choose the best ML algorithm after training (data for chromosome IV was not considered for model selection). NFRs used for learning were those located by nucleR in between two well-positioned nucleosomes and placed close to TSS or TTS. Non-NFR regions were those occupied by nucleosomes, as non-NFR are longer than NFR, we randomly removed points from non-NFR regions to obtain a balanced data set of ones (points in an NFR) and zeros (points in a non-NFR region). The descriptive features are the values of the deformation energy and TFBS density 268 bps around each point (the mean width of NFRs is 268 in our data).

Exhaustive analysis reveals the Extra-Trees classifier as the optimum predictor (converged values were obtained at around 500 trees; see Suppl. Figure S2) for which AUC (area under the ROC curve) around 0.93 was obtained in the validation subset. The final model was revalidated using Chromosome IV data that were not considered at any point of the development of the ML algorithm. ML models were built using the Python library scikit-learn (v 0.20.3) (Pedregosa et al. 2011).

RESULTS AND DISCUSSION

Nucleosome positioning at the gene body is determined by periodicity. We computed the autocorrelation coefficient from the nucleosome coverage in our MNase-seq experiments for different periods (T ; see eq. 1-2), finding a clear peak at 165 bp (the average nucleosome repeat length in yeast, (Ocampo et al. 2016); see Suppl. Figure S3), which marks the distance (DFI, see Methods) between the +1 and the -last nucleosomes, but not the distance between the TSS and TTS where no phasing bias is detected (see Figure 1) due probably to the different distances between the TSS and the +1 nucleosome and/or between the TTS and the -last. Taking the positions of +1 and -last nucleosomes as emitting sites of a distance decaying signal (see Methods, eq. 4-6) we can reproduce extremely well the nucleosome architecture within the gene (Figure 2). Notably, we can also distinguish between phased genes, where the +1 and -last signals add up to define clear and periodic nucleosome patterns, and unphased genes, where signals can partially cancel out in the middle of the gene, leading to diffuse nucleosome patterns (see examples marked with arrows in Figure 2A, and profiles in Figure 2B).

To confirm that periodicity guides the positioning of the nucleosomes in the +1 to -last segment we selected 4 phased and 4 control, not-phased, genes (see Methods, Suppl. Table S1) and inserted an innocuous 81-nt oligo (see *Methods*) which should destroy phasing of the first group, affecting little already not-phased genes. As expected from our model, change in phasing in the first group of genes leads to an important loss of periodicity and to fuzzier nucleosome arrangements, while small changes are found in the control genes (Table S3, Figure 3, Suppl. Figures S4 and S5). Interestingly, the placements of the +1 and -last nucleosomes are not altered by adding the 81-nt fragment (Figure 3) and no change in nucleosome structure is found in neighboring genes, in perfect agreement with our model. In summary, our synthetic biology experiments fully confirm that nucleosomes arrange in a local and periodic way from strong eviction signals at the NFR located at the TSS and TTS and simple periodicity considerations in the intragenic region.

NFRs are characterized by unique physical properties and by the high density of protein-recognition sequences. Analysis of the entire yeast genome illustrates NFRs at the TSS and TTS of the gene (TSS-NFR, TTS-NFR). The latter is present in both tandem and convergent genes, showing that the TTS-NFR is not a duplication of a neighboring TSS (see Suppl. Figure S6). Interestingly, both NFRs correspond to regions where the elastic energy associated to wrap the DNA around the histone core is unusually large and where there is a large density of potential TFBS (see Methods and Figure 4A and 4B). These two signals can be used to train a ML classifier (see above), which shows a large ability to detect NFRs, as shown in the area under the ROC curve above 0.94 in the validation subset (see Figure 4C and Methods for details). Predictions on TSS-NFRs and TTS-NFRs for chr I show sensitivity and precision of 0.66 and 0.57, respectively. These values are improved taking into account also other NFRs that the classifier correctly predicts (sensitivity 0.76 and precision 0.58, AUC 0.77) but that are not associated to TSS or TTS, indicating that the method has predictive power out of the specific characteristics of the training set. Moreover, the predictor also has good accuracy on fragile nucleosomes (Suppl. Figure S7) as defined in (Kubik et al. 2015). In summary, DNA sequence guides the location of regions with a tendency to be depleted in nucleosomes by a double mechanism: i) by creating regions where effector proteins compete with nucleosomes for binding the DNA and ii) by defining stiff regions, where it is difficult to wrap the DNA around the histone octamer. It is tempting to believe that both factors are coupled, as probably evolution guided the placement of TFBS in regions, which were nucleosome free. In any case, once the NFR is broadly defined, the placement of the +1 and -last nucleosome and accordingly the architecture of the entire nucleosome fiber (see Figure 2) is likely to be defined based on the need to preserve the periodicity to place the maximum number of nucleosomes in the +1 to -last segment.

Nucleosome architecture and gene expression are coupled in a complex way. As described before (Weiner et al. 2010)(Chereji and Morozov 2015), (Deniz et al., 2011, 2016) genes with wide TSS-NFRs tend to be transcriptionally active (Figure 5A). This finding has been explained (Shivaswamy et al. 2008; Huebert et al. 2012; Deniz et al. 2016) considering that wide NFRs should favor the recruitment of effector proteins and RNA polymerase,

suggesting a causal relationship: *widening of the NFR* \rightarrow *increase in transcription activity*. The same genome-wide analysis shows that, transcriptionally active genes have more periodic nucleosome distributions in the gene body than the inactive ones (Figure 5B). This finding suggests another causal relationship: *more periodic nucleosomes* \rightarrow *greater transcriptional activity*, which agree with previous suggestions that more compact nucleosomal arrays implies higher transcription rate (Vaillant et al., 2010; Deniz et al., 2016)

We investigated the causal relationships suggested above (Figure 5): *wide NFR* \rightarrow *more active genes* and *phased nucleosomes* \rightarrow *more active genes* by different approaches. Firstly, we repeated nucleosome positioning experiments, but treating previously the cells with 1,10-Phenanthroline, a metal chelator that avoids polymerase progression (McClure et al. 1978) without affecting binding to the core promoter (Grigull et al. 2004; Kim et al. 2010). Results summarized in Figure 5C show that inhibition of transcription by 1,10-Phenanthroline leads to wider TSS-NFRs (around 45% of the genes show increases above 10 bps), due in general to the displacement upstream of the -1 nucleosome (see Figure 5D). Notice that these changes are not due to differences in MNase digestion, as shown by the comparable length of the sequenced fragments (Suppl. Figure S8). This suggests that when RNA polymerase binds to the core promoter sliding of the -1 nucleosome happens and the associated NFR becomes wider. In untreated cells, once polymerase displaces along the gene body, the density of RNA polymerase at core promoter decreases and the -1 nucleosome partially recovers its normal placement, narrowing the NFR. In the presence of 1,10-Phenanthroline the core promoter is saturated by RNA polymerase and accordingly NFR widens. This implies that TSS-NFRs are more plastic than expected and can adapt its width to the presence of effector proteins, specifically RNA polymerase. This plasticity is observed for example when cells are subjected to stress and the osmo-responsive genes experience a widening of their TSS due to a shift of the +1 nucleosome in presence of HOG1 (Nadal-Ribelles et al. 2012). In other words, the causal relationship between NFR width and gene activity seems to go in the direction that makes RNA polymerase binding responsible for NFR widening in active genes, i.e. *gene activity* \rightarrow *wide NFRs*, not the reverse causal relationship.

The addition of 1,10-Phenanthroline produces significant changes in nucleosome architecture also in the gene body, leading to fuzzier and less periodic nucleosome positions (Figures 5E-F). As the effect of the drug is to avoid polymerase migration from the TSS, we can conclude that the lack of RNA-polymerase in the gene body leads to a loss of periodicity and to fuzzier nucleosomes. This agrees with a causal relationship: *gene activity* \rightarrow *periodic nucleosomes*, and is contrary to the idea that periodic nucleosomes triggers gene activity. This causal relationship is also confirmed by the negligible changes in the level of expression found in phased genes when an inert 81-not long oligo is inserted (Suppl. Figure S9), which contrast with the significant increase in nucleosome fuzziness. Thus, our results strongly suggest that RNA polymerase activity generates temporary eviction signals that help to organize the nucleosome fiber in a more periodic manner, but having a well-ordered or a fuzzy nucleosome string in the gene body do not significantly alter gene activity.

CONCLUSIONS

Nucleosome positioning in the gene body can be predicted with good accuracy by assuming that two well-positioned nucleosomes emit a periodic signal whose intensity is decaying with sequence. Phased genes (i.e. those whose $+1 \leftrightarrow \text{last}$ distance is multiple of 165) have periodic nucleosome signals, while unphased genes (and at a lower extend non-phased genes) tend to have fuzzy nucleosomes in the middle of the gene body. Change in the $+1 \leftrightarrow \text{last}$ distance leads to changes in nucleosome periodicity fully predictable by the theory. Very interestingly, the placement of the $+1$ and $-\text{last}$ nucleosomes are defined by the vicinity of NFRs, i.e., segments of DNA depleted of nucleosomes. Such NFRs are characterized by DNA physical properties that disfavor nucleosome formation and by high density of TFBS. The combined signal is strong enough to allow training of ML methods, which predict NFRs with good accuracy.

Wide TSS NFRs and periodic nucleosome architectures in the gene body signal transcriptionally active genes. However, our results strongly suggest that NFR width might be a consequence of the binding of RNA polymerase rather than the cause of it. Furthermore, experiments with inhibition of RNA pol progression and synthetic experiments where phasing was artificially altered did not reveal any significant change in transcriptional activity, related to the level or order in the nucleosome array in the gene body suggesting that a periodic nucleosome string might be a consequence and not a cause of RNA polymerase activity. This observation is consistent with recent results (Vasseur et al. 2016) showing that the genes that exhibit comparatively rapid phasing of nucleosomes over gene bodies after replication are relatively highly transcribed. In summary our results suggest a causal relationship in the correlation between gene activity and nucleosome architecture, as the influence of gene activity in the placement of nucleosomes seems to be larger than the influence of nucleosome architecture in the activity of the gene.

Acknowledgements

We thank the Spanish Ministry of Science [Plan Nacional], Catalan SGR, the Instituto Nacional de Bioinformática; European Research Council (ERC SimDNA), European Union's Horizon 2020 research and innovation program [676556], Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) and H2020 BioExcel project (all awarded to M.O.); MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona). MO is an ICREA Research Fellow.

References

- Breiman L. 2001. Random Forests. *Machine Learning* **45**: 5-32.
- Buitrago D, Codó L, Illa R, de Jorge P, Battistini F, Flores O, Bayarri G, Del Pino M, Heath S, Hospital A et al. In press. Nucleosome Dynamics: A new tool for the dynamic analysis of nucleosome positioning. *Nucleic Acids Res.* In Press, 2019
- Chereji RV, Clark DJ. 2018. Major Determinants of Nucleosome Positioning. *Biophysical journal* doi:10.1016/j.bpj.2018.03.015.
- Chereji RV, Morozov AV. 2015. Functional roles of nucleosome stability and dynamics. *Brief Funct Genomics* **14**: 50-60.
- Clark DJ. 2010. Nucleosome positioning, nucleosome spacing and the nucleosome code. *J Biomol Struct Dyn* **27**: 781-793.
- Dans PD, Balaceanu A, Pasi M, Patelli AS, Petkeviciute D, Walther J, Hospital A, Lavery R, Maddocks JH, Orozco M. In press. The physical properties of B-DNA beyond Calladine-Dickerson rules. *Nucleic Acids Research*.
- Deniz O, Flores O, Aldea M, Soler-Lopez M, Orozco M. 2016. Nucleosome architecture throughout the cell cycle. *Scientific reports* **6**: 19729.
- Deniz O, Flores O, Battistini F, Perez A, Soler-Lopez M, Orozco M. 2011. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics* **12**: 489.
- Flores O, Orozco M. 2011. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**: 2149-2150.
- Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. *Machine Learning* **63**: 3-42.
- Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR. 2004. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol* **24**: 5534-5547.
- Huebert DJ, Kuan PF, Keles S, Gasch AP. 2012. Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. *Mol Cell Biol* **32**: 1645-1653.
- Hughes AL, Jin Y, Rando OJ, Struhl K. 2012. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. *Molecular cell* **48**: 5-15.
- Izzo A, Kamieniarz K, Schneider R. 2008. The histone H1 family: specific members, specific functions? *Biol Chem* **389**: 333-343.
- Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**: 161-172.
- Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362-366.
- Kim TS, Liu CL, Yassour M, Holik J, Friedman N, Buratowski S, Rando OJ. 2010. RNA polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast. *Genome Biol* **11**: R75.
- Krietenstein N, Wal M, Watanabe S, Park B, Peterson CL, Pugh BF, Korber P. 2016. Genomic Nucleosome Organization Reconstituted with Pure Proteins. *Cell* **167**: 709-721 e712.
- Kubik S, Bruzzzone MJ, Jacquet P, Falcone J-L, Rougemont J, Shore D. 2015. Nucleosome Stability Distinguishes Two Different Promoter Types at All Protein-Coding Genes in Yeast. *Molecular Cell* **60**: 422-434.

- Kubik S, Bruzzone MJ, Challal D, Dreos R, Mattarocci S, Bucher P, Libri D, Shore D. 2019. Opposing chromatin remodelers control transcription initiation frequency and start site selection. *Nat Struct Mol Biol* **26**: 744-754.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Lieleg C, Krietenstein N, Walker M, Korber P. 2015. Nucleosome positioning in yeasts: methods, maps, and mechanisms. *Chromosoma* **124**: 131-151.
- Lorch Y, Maier-Davis B, Kornberg RD. 2014. Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes Dev* **28**: 2492-2497.
- Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008a. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**: 1073-1083.
- Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC et al. 2008b. Nucleosome organization in the Drosophila genome. *Nature* **453**: 358-362.
- McClure WR, Cech CL, Johnston DE. 1978. A steady state assay for the RNA polymerase initiation reaction. *J Biol Chem* **253**: 8941-8948.
- Nadal-Ribelles M, Conde N, Flores O, González-Vallinas J, Eyraas E, Orozco M, de Nadal E, Posas F. 2012. Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling. *Genome Biol* **13**: R106.
- Nocetti N, Whitehouse I. 2016. Nucleosome repositioning underlies dynamic gene expression. *Genes Dev* **30**: 660-672.
- Ocampo J, Chereji RV, Eriksson PR, Clark DJ. 2016. The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. *Nucleic Acids Res* **44**: 4625-4635.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**: 2825-2830.
- Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. *Nature* **423**: 145-150.
- Schlenstedt G, Hurt E, Doye V, Silver PA. 1993. Reconstitution of nuclear protein transport with semi-intact yeast cells. *J Cell Biol* **123**: 785-798.
- Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* **6**: e65.
- Storici F, Resnick MA. 2006. The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods in enzymology* **409**: 329-345.
- Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**: 267-273.
- Suter B, Schnappauf G, Thoma F. 2000. Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res* **28**: 4083-4089.
- Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res* **20**: 59-67.
- Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* **474**: 516-520.

- Vasseur P, Tonazzini S, Ziane R, Camasses A, Rando OJ, Radman-Livaja M. 2016. Dynamics of Nucleosome Positioning Maturation following Genomic Replication. *Cell reports* **16**: 2651-2665.
- Walther J, Dans PD, Balaceanu A, Hospital A, Bayarri G, Orozco M. In revision. A multi-modal coarse-grain model of DNA flexibility mappable to the atomistic level. *Nucleic Acids Research*.
- Wan J, Lin J, Zack DJ, Qian J. 2009. Relating periodicity of nucleosome organization and gene regulation. *Bioinformatics* **25**: 1782-1788.
- Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N. 2010. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res* **20**: 90-100.
- Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**: 238-241.
- Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**: 626-630.
- Zhang Z, Pugh BF. 2011. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* **144**: 175-186.

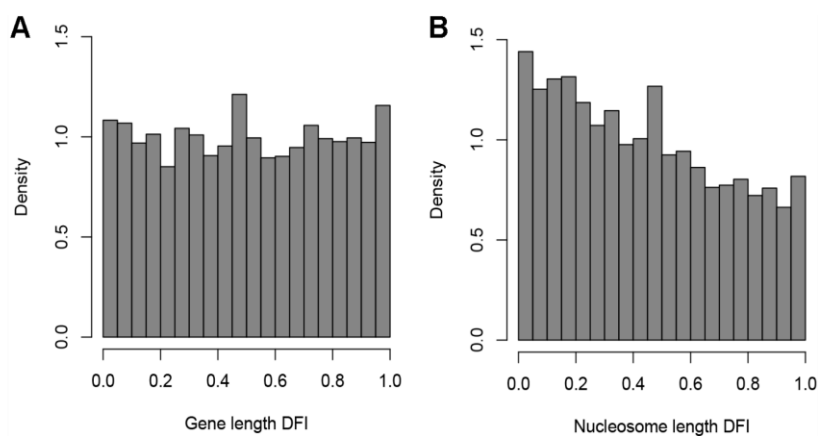


Figure 1. Distance from integer (DFI, see Methods) score computed on (A) the gene length (distance between TSS and TTS) and (B) the nucleosome length (distance between +1 and -last nucleosome). DFI is normalized (between 0 and 1) dividing by $T/2$ and taking the absolute value.

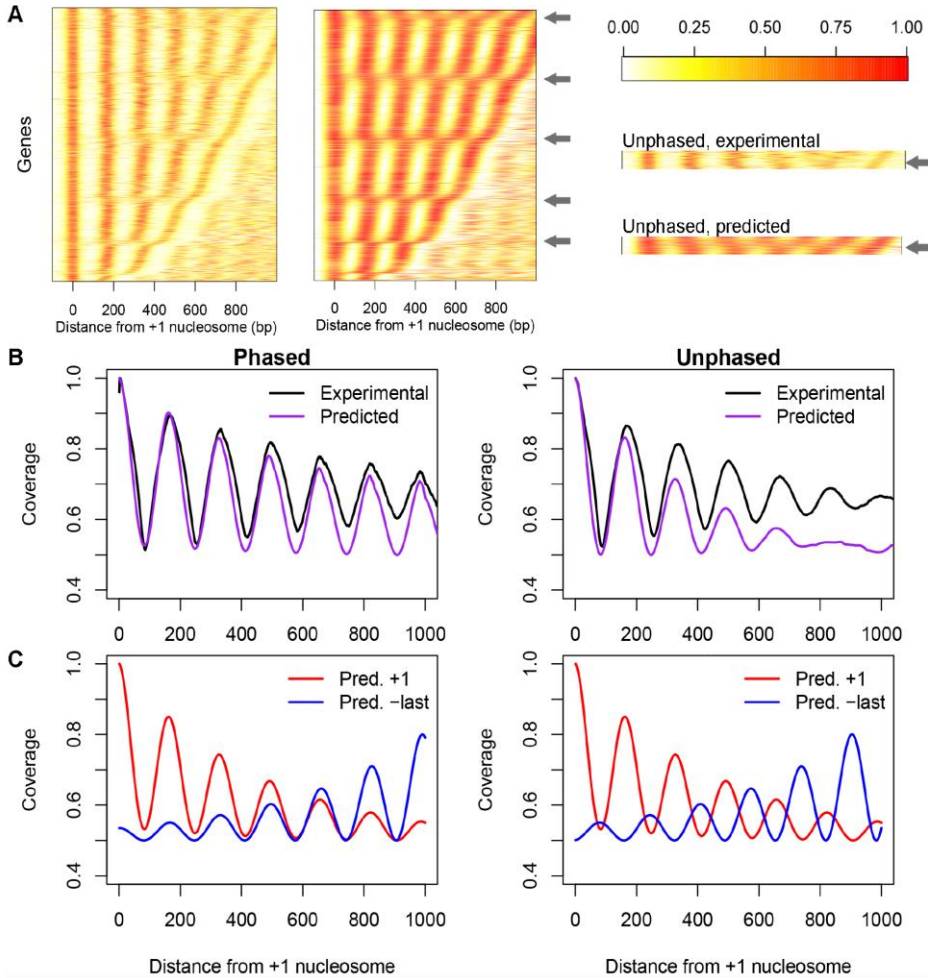


Figure 2. Signal decay model of nucleosome positioning. (A) Experimental (left panel) and predicted (right panel) nucleosome coverage for each gene, with respect to +1 nucleosome. Genes are sorted by the distance between +1 and -last nucleosomes. Colour scale corresponds to normalized nucleosome coverage, from 1 (red) to 0 (yellow). (B) Nucleosome coverage, experimental (black) and predicted (purple, see Methods eq. 6) from +1 nucleosome, averaged across all genes. Genes are split into phased (left) or unphased (right) based on $DFI < 10$ and $DFI > 40$, respectively. (C) Signals from +1 (red) and -last nucleosomes (blue) to predict the experimental coverage of phased (left) and unphased (right) genes (see Methods eq. 4-5).

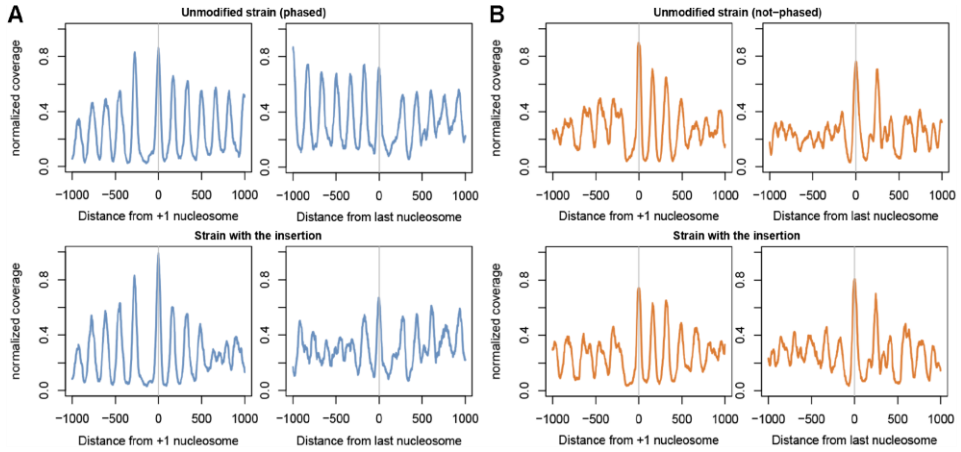


Figure 3. Nucleosome coverage for the selected genes in the unmodified strain (top panel) and the strain with the 81-nt insertion (bottom panel). (A) Average of the four genes phased in the unmodified strain (UBX5, CKB2, PPT1 and TRP4) and (B) the four genes not-phased in the unmodified strain (BSP1, DGK1, SLM3 and PAN5).

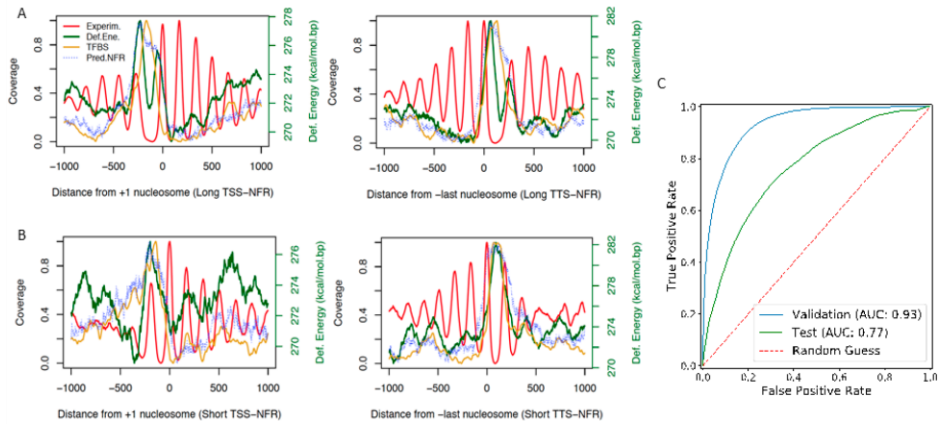


Figure 4. Average nucleosome coverage (red), TFBS density (yellow), deformation energy (green) and NFR prediction (blue) around +1 and -last nucleosomes for (A) long NFRs (wider than 215 bps) and (B) short NFRs (shorter than 215 bps). (C) Results of Extra-Trees classifier for the prediction of NFRs from deformation energy and TFBS density in the validation subset and for the entire chromosome I.

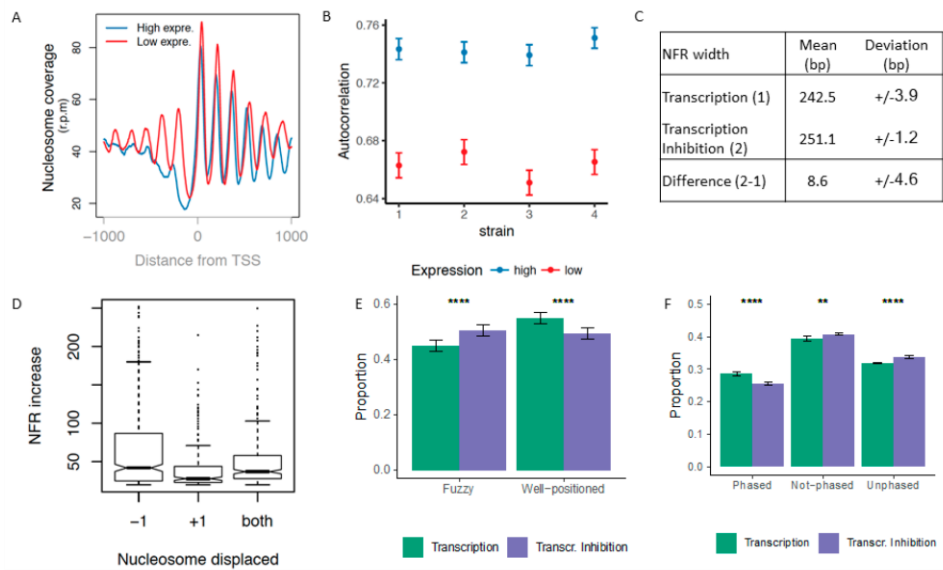


Figure 5. Effect of transcription on nucleosome positioning. (A) Nucleosome coverage around TSS for highly (blue) and lowly (red) expressed genes. (B) Autocorrelation score for highly (blue) and lowly (red) expressed genes. (C) Change in TSS-NFR width (-1 to +1 nucleosome distance) upon transcription inhibition in the presence of 1,10-Phenantroline. (D) TSS-NFR increase for genes displacing -1, +1 or both nucleosomes upon transcription inhibition. (E) Change in the proportion of Fuzzy and Well-positioned nucleosomes upon transcription inhibition, with bars indicating relative standard error. (F) Change in the proportion of phased, not-phased and unphased genes upon transcription inhibition, with bars indicating relative standard error.

Bibliography for Chapter 4

- [1] W. K. M. Lai and B. F. Pugh, "Understanding nucleosome dynamics and their links to gene expression and DNA replication," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 9, pp. 548–562, May 2017.
- [2] C. Jiang and B. F. Pugh, "Nucleosome positioning and gene regulation: advances through genomics," *Nat. Rev. Genet.*, vol. 10, no. 3, pp. 161–172, Mar. 2009.
- [3] A. Weiner, A. Hughes, M. Yassour, O. J. Rando, and N. Friedman, "High-resolution nucleosome mapping reveals transcription-dependent promoter packaging," *Genome Res.*, vol. 20, no. 1, pp. 90–100, Jan. 2010.
- [4] R. V. Chereji and A. V. Morozov, "Functional roles of nucleosome stability and dynamics," *Brief. Funct. Genomics*, vol. 14, no. 1, pp. 50–60, Jan. 2015.
- [5] Ö. Deniz, O. Flores, F. Battistini, A. Pérez, M. Soler-López, and M. Orozco, "Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast," *BMC Genomics*, vol. 12, no. 1, Dec. 2011.
- [6] Ö. Deniz, O. Flores, M. Aldea, M. Soler-López, and M. Orozco, "Nucleosome architecture throughout the cell cycle," *Sci. Rep.*, vol. 6, p. 19729, Jan. 2016.
- [7] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa, and C. Thermes, "Multi-scale coding of genomic information: From DNA sequence to genome structure and function," *Phys. Rep.*, vol. 498, no. 2–3, pp. 45–188, Feb. 2011.
- [8] R. V. Chereji, S. Ramachandran, T. D. Bryson, and S. Henikoff, "Precise genome-wide mapping of single nucleosomes and linkers in vivo," *Genome Biol.*, vol. 19, no. 1, Dec. 2018.
- [9] J. Grigull, S. Mnaimneh, J. Pootoolal, M. D. Robinson, and T. R. Hughes, "Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors," *Mol. Cell. Biol.*, vol. 24, no. 12, pp. 5534–5547, Jun. 2004.
- [10] T. S. Kim *et al.*, "RNA polymerase mapping during stress responses reveals widespread nonproductive transcription in yeast," *Genome Biol.*, vol. 11, no. 7, p. R75, 2010.
- [11] W. R. McClure, C. L. Cech, and D. E. Johnston, "A steady state assay for the RNA polymerase initiation reaction," *J. Biol. Chem.*, vol. 253, no. 24, pp. 8941–8948, Dec. 1978.
- [12] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

Chapter 5 . Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning

Nucleosomes are the main unit of eukaryotic chromatin, modulating the accessibility of DNA to effector proteins. The nucleosome architecture is then related to gene regulation, DNA replication and other cellular processes [1], [2]. Accordingly, determination of the arrangement of nucleosomes in the cell is crucial to gain a function view on the chromatin structure. Our group previously developed an algorithm, nucleR [3], for nucleosome positioning from experimental MNase-seq data (see details in Chapter 2) that is one of the most widely used tools to define average nucleosome configuration in living cells. Unfortunately, despite its power nucleR presents a series of shortcomings that lead us to develop a more universal and flexible tool. First, due to the noisy nature of experimental data such as MNase-seq, it is not easy to compare the results obtained between two experimental conditions to study the dynamics of nucleosome organization when changing cellular conditions. Second, analyses are not automatized, nor standardized or FAIR-certified. Third, it is not easy to directly integrate the results from nucleR with other features obtained from several techniques (ChIP-seq, RNA-seq, etc.). Finally, other experimental protocols (different to MNase-seq) for nucleosome mapping have emerged in recent years and it is not clear whether nucleR can be applied to this type of data. Therefore, in this chapter, I present the

theoretical and methodological developments we performed to cope with the above-mentioned user demands.

First, we developed an algorithm for the differential analysis of two MNase-seq experiments, **NucDyn**. It uses directly the mapped reads, allowing the detection of changes that occur even in a small percentage of the cells in the population. The method, very robust and fast, vastly outperforms other available software, such as Danpos [4] or Dimnp [5]. NucDyn employs a dynamic programming algorithm and statistical metrics detecting, even in noisy experiments, changes in nucleosome architecture (see **Figure 5.1**).

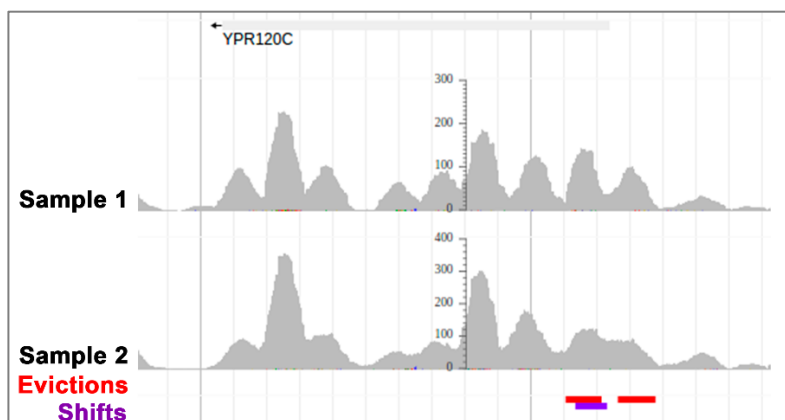


Figure 5.1. Comparison of two nucleosome profiles obtained from MNase-seq (Sample 1 and Sample 2) using NucDyn (evictions and shifts identified).

To gain a more quantitative description of changes in nucleosome arrangements, we developed a series of methods and metrics (see **Figure 5.2**[Error! Reference source not found.](#)):

- i. **Periodicity:** the *autocorrelation* and *phasing* scores (see Chapter 4) to study the periodicity of nucleosomes along the gene body can be computed from the nucleR results. Also, nucleosome occupancy profiles can be predicted based on signal propagation theory using two opposing emitting signals from the +1 and the -last nucleosomes.
- ii. **TSS-classification:** each gene can be characterized by the nucleosome free region (NFR) around its promoter (open (o), closed (c) or missing -1 or +1

nucleosome) and the degree of localization of the +1 (downstream the TSS) and -1 (upstream the TSS) nucleosomes (fuzzy or well-positioned).

- iii. **Nucleosome stiffness:** sliding propensity of a nucleosome can be estimated from the variability in nucleosome position among the cell population. A Gaussian curve is fitted to the dyad distribution for each nucleosome and the estimated standard deviation is used to derive the apparent stiffness by an elastic approximation.
- iv. **NFR detection:** from the nucleR results we can detect NFRs, excluding low mappability regions. These loci are typically related to regulatory elements such as transcription factor binding sites or replication origins.

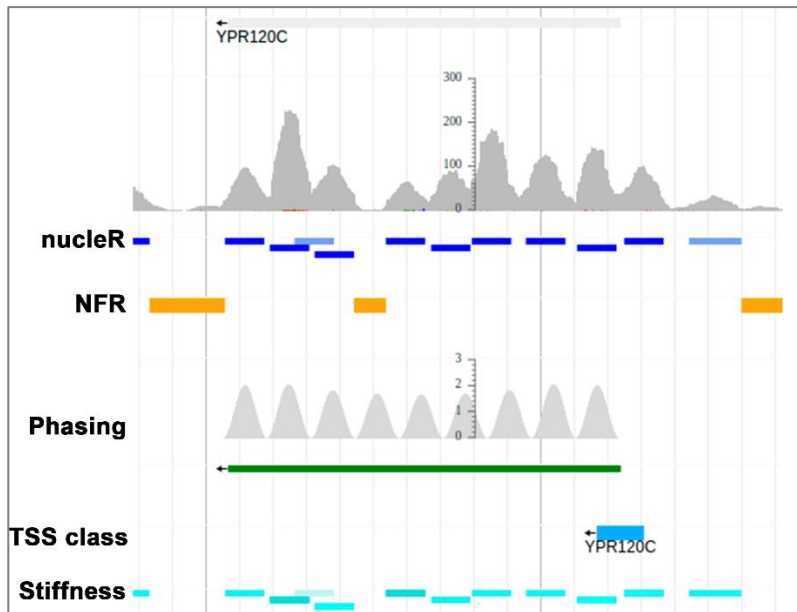


Figure 5.2. Example of results of nucleR and other nucleosome-related analyses using MNase-seq data (coverage in grey).

We have integrated these tools, written in R and available in our GitHub repository (<https://github.com/nucleosome-dynamics>), into a package called Nucleosome Dynamics (available under the Apache 2.0 License). Besides directly running the R source code, it can be executed under several distribution models: as a software container (Docker and Singularity containers are available) that includes all the

dependencies required, or as a web tool in the MuGVRE workspace [6] or in a Galaxy server [7]. In the web implementations, the user can either upload files containing the mapped reads (typically BAM files) or upload the sequencing files (FASTQ) and use one of the aligners available to map the reads to the reference genome. Then, the user can execute all (or some) of the tools available in Nucleosome Dynamics and monitor the status of the calculations from the web. The documentation, tutorials and access to the different distribution options are summarized in the web <http://mmb.irbbarcelona.org/NucleosomeDynamics/>.

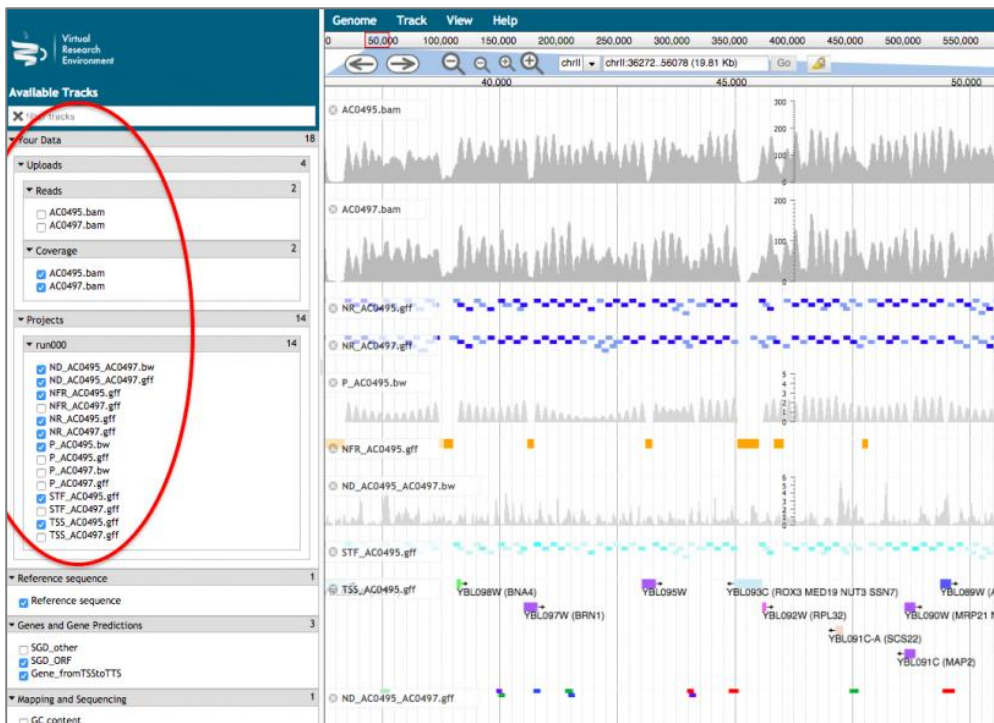


Figure 5.3. Example of visualization of Nucleosome Dynamics results in the MuGVRE.

The MuGVRE allows the representation of the results from nucleR, NucDyn and all nucleosome related analyses in an integrated genome browser (see **Figure 5.3**), where additional genomic data can be contrasted and jointly analyzed with the nucleosome information. Furthermore, summary statistics from every one of our tools are automatically generated for every gene as well as some results at the genome-wide level.

Finally, we explored the performance of our algorithms to analyze data obtained with recent chemical cleavage methods for nucleosome positioning, that have been proposed aiming to remove the sequence bias and the effect of digestion level of MNase [8], [9]. However, these methods have another limitation for their broad use, since this technique requires genetic engineering replacing the endogenous histone H4 or H3 by a mutated version. Nonetheless, we have demonstrated with some published data from [9] that Nucleosome Dynamics can also be applied to analyze nucleosome profiles from chemical cleavage data.

We have shown the usefulness of Nucleosome Dynamics in several experimental settings (changes throughout the cell cycle, along the yeast metabolic cycle and in response to different carbon sources) where it allowed to correlate changes in nucleosome organization with differential gene activity. The package is presented in the publication *Nucleosome Dynamics: A new tool for the dynamic analysis of nucleosome positioning* attached in the following pages.

Publication:

Diana Buitrago^{*}, Laia Codó^{*}, Ricard Illa, Pau de Jorge, Federica Battistini, Oscar Flores, Genis Bayarri, Romina Royo, Marc Del Pino, Simon Heath, Adam Hospital, Josep Lluís Gelpí, Isabelle Brun Heath and Modesto Orozco. (2019). Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning. Nucleic Acids Research, gkz759. <https://doi.org/10.1093/nar/gkz759>.

Supplementary material for this article can be found in the **Annex IV**.

^{*} Equally contributing authors

Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning

Diana Buitrago^{1,†}, Laia Codó^{2,†}, Ricard Illa¹, Pau de Jorge¹, Federica Battistini¹, Oscar Flores¹, Genis Bayarri¹, Romina Royo², Marc Del Pino², Simon Heath³, Adam Hospital¹, Josep Lluís Gelpí^{2,4}, Isabelle Brun Heath¹ and Modesto Orozco^{1,4,*}

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldri Reixac 10, Barcelona 08028, Spain, ²Barcelona Supercomputing Center (BSC), Jordi Girona 31, Barcelona 08028, Spain, ³Centro Nacional de Análisis Genómico (CNAG-CRG), Centre de Regulació Genómica (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain and ⁴Departament de Bioquímica i Biomedicina, Facultat de Biologia, Universitat de Barcelona, Avda Diagonal 647, Barcelona 08028, Spain

Received June 03, 2019; Revised July 23, 2019; Editorial Decision August 06, 2019; Accepted August 22, 2019

ABSTRACT

We present **Nucleosome Dynamics**, a suite of programs integrated into a virtual research environment and created to define nucleosome architecture and dynamics from noisy experimental data. The package allows both the definition of nucleosome architectures and the detection of changes in nucleosomal organization due to changes in cellular conditions. Results are displayed in the context of genomic information thanks to different visualizers and browsers, allowing the user a holistic, multidimensional view of the genome/transcriptome. The package shows good performance for both locating equilibrium nucleosome architecture and nucleosome dynamics and provides abundant useful information in several test cases, where experimental data on nucleosome position (and for some cases expression level) have been collected for cells under different external conditions (cell cycle phase, yeast metabolic cycle progression, changes in nutrients or difference in MNase digestion level). **Nucleosome Dynamics** is a free software and is provided under several distribution models.

INTRODUCTION

Eukaryotic chromatin is organized in a hierarchical manner, where the basic structural units are repetitive elements named nucleosomes. Each of them is defined by around 147 base pairs of DNA wrapped around a protein octamer, the histones. The position of the nucleosomes in the cell is

not random and recurrent patterns have been detected in cell populations (1–3), indicating a maintenance of the nucleosome architecture which seems to be crucial for a correct regulation of genome activity (4). The protein octamer serves as an anchoring point for proteins recognizing histone epigenetic signals, while unwrapped DNA is targeted by transcription factors and enhancers (5,6). Thus, nucleosomes shifting due to alterations in the sequence (7), DNA methylation (8) or the action of chromatin remodelers (5,9–13) can result in dramatic changes in gene expression. Characterizing such changes is crucial for the understanding of the connection between chromatin structure and genome functionality (6).

Experimental determination of nucleosome positioning is typically performed by treating a group of cells (in the range 10^6 – 10^9) with enzymes acting on nucleosome-free DNA. ATAC-seq (14) uses a hyperactive transposase for tagging nucleosome-free DNA segments for sequencing (the linkers). MNase-seq, the most widely used technique for nucleosome localization, uses Micrococcal nuclease to degrade linker DNA preserving the DNA segments wrapped in the nucleosomes, which are then sequenced. Both MNase-seq and ATAC-seq, after filtering nucleosomal reads by size (14), provide at the end the same type of information: DNA reads that need to be grouped into individual nucleosomes using a variety of computational approaches (15–18), which in all cases suffer from the intrinsic dispersion in read coverage. The resulting nucleosome maps show well defined depleted regions (the nucleosome free regions, NFR), well-positioned (W) nucleosomes, and a large number of ‘fuzzy’ (F) nucleosomes giving partial protection signals longer than 147 bps (19). Fuzzy positioning signals are the result of nucleosomes not being in exactly the same

*To whom correspondence should be addressed. Tel: +34 93 40 37156; Fax: +34 93 40 37157; Email: modesto.orozco@irbbarcelona.org

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 Nucleic Acids Research, 2019

genomic position in the cell population, and are intrinsically difficult to annotate by any nucleosome calling algorithms (15,17). While it is known (20,21) that this technique can be affected by MNase concentration and sequence-preference biases that affect the detection of the so called 'fragile' nucleosomes, it is still the most widely used to detect nucleosome positioning for its versatility and accuracy. In 2012, Brogaard *et al.* developed a chemical cleavage method that provides a very accurate positioning of nucleosomes (22). However, this technique requires to do genetic engineering replacing the endogenous histone H4 (or H3 (23)) by a mutated version, therefore restricting its use (24–27). Moreover, it has been shown that the MNase sequence bias can be corrected using digested naked DNA as baseline (20,21), obtaining more pronounced nucleosome coverage peaks.

The noisy nature of experimental data such as MNase-seq, makes very difficult to compare nucleosome architecture in two samples, as the signal is masked by the intrinsic fuzziness of the maps. Methods available such as DAN-POS and Dimnp (15,28) can detect only a limited number of changes affecting large percentage of the cells, as they work at the level of the fragment coverage, missing the opportunity to work with the raw data: the fragments themselves.

We present here Nucleosome Dynamics, a complete virtual framework to characterize the structure and dynamics of nucleosome architectures. The package consists of two main blocks: an improved version of our nucleR algorithm for nucleosome location (17), and NucDyn, an algorithm specifically created to detect changes (shifts, evictions and insertions) in nucleosome architectures based on the direct processing of raw data (the sequencing reads) obtained from pairs of experiments. The Nucleosome Dynamics package (available under the Apache 2.0 License) can be installed from the source code, obtained from BioConductor (29), or run as a web tool hosted by the MuGVRE workspace (30) as well as in a Galaxy server (31), where additional analysis algorithms, browsers and visualization tools are included.

MATERIALS AND METHODS

Package overview

The input data for Nucleosome Dynamics is one or several files containing sequence reads aligned to the reference genome and stored in BAM format. The user can select (see Figure 1) between: (i) processing a single file to define the consensus nucleosome architecture using an extended version of nucleR (17) or (ii) detecting changes in nucleosome distribution between two experiments, by comparing pairs of mapped sequence files using the newly developed NucDyn module. For a complete description of the parameters of the available functionalities, see Supplementary Table S1. In the MuGVRE implementation (see Table 1), the user has access to a wide range of analysis and visualization tools to characterize the nucleosome patterns, their changes across different conditions, and to put all the data in the context of other information mapped to the genome (genome structure, expression, epigenetic signals, etc). We have evaluated the performance of nucleR and NucDyn generating synthetic nucleosome maps and have tested their descriptive power using publicly available nucleosome positioning experimental data.

Single experiment analysis

Nucleosome positioning and coverage. BAM files from a single MNase-seq experiment are processed to define nucleosome coverage, which can be directly visualized using a genome browser (Figure 2A) or processed to obtain nucleosome positions. Accordingly, following signal theory the read coverage is described as a combination of periodic waves, which are then subjected to Fast Fourier Transformation (FFT) to remove the high frequency components responsible for the noise (see Supplementary Figure S1). The parameters for FFT filtering can be adjusted taking into account the nucleosome repeat length and noise level of each organism and cell type (see Supplementary Table S1). Clean profiles are processed to annotate the nucleosome dyads (located at the local maxima of the distributions). Putative nucleosomes are then scored based on the shape of the associated peaks (see Supplementary Figure S1). Those leading to sharp signals are labelled as well-positioned (W) nucleosomes (high localization score), while flat peaks are labelled as 'fuzzy' (F) nucleosomes (low localization score). Once all nucleosomes are localized, the software analyses the nucleosome architecture (see Supplementary Figure S2A) around the transcription start sites (TSS) and classifies the nucleosome architecture for each gene based on (19): the extension of the nucleosome free region (NFR) around the core promoter (open (o), closed (c) or missing –1 or +1 nucleosome) and the degree of localization of the +1 (downstream the TSS) and –1 (upstream the TSS) nucleosomes (see Supplementary Figure S2A). Data are presented at the individual gene level as well as summarized at global level (Figure 2). Nucleosome Dynamics performs a global detection of all NFRs, as these regions usually are the main recognition sites for effector proteins, and well-defined and extended NFRs typically signal active regions in the genome.

Periodicity at coding regions. The software evaluates the periodicity in the nucleosome pattern inside the genes, following signal propagation theory from two 'emitting sites' located at well-positioned nucleosomes. The first signal comes from the 5' end of the gene (the +1 nucleosome located just downstream the TSS) and the second from the 3' end of the gene (the –last nucleosome located just upstream the transcription termination site; TTS). We assume that both signals proceed in opposite directions (from +1 to –last nucleosome) following an exponential decay periodic function (32). We found out that when the +1 and –last originated waves are in phase the signals sum up and nucleosomes are well located inside the gene body, while when they are in antiphase the signals cancel out and the gene typically shows fuzzy nucleosomes. The periodicity (T) of the signal is obtained by maximizing the autocorrelation function (Equation 1: see an example in Supplementary Figure S2):

$$R(T) = \int_{X_2}^{X_1} I(x) * I(x - T) dx \quad (1)$$

where X_1 and X_2 are the intervals of the window, $I(x)$ stands for the coverage. This value will be dependent on the nucleosome repeat length of each species and cell type (see Supplementary Table S2 for suggested T in different cell types).

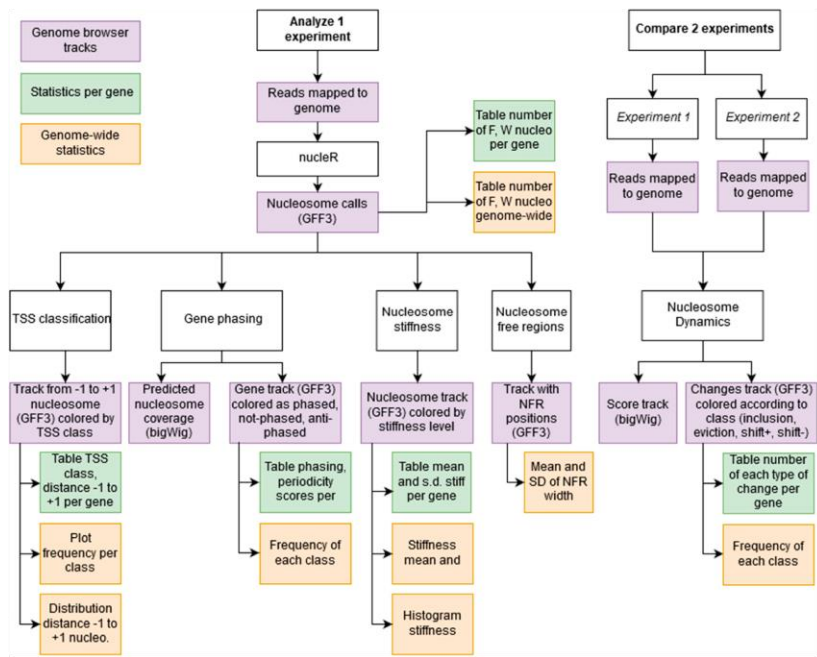


Figure 1. Analysis pipeline for Nucleosome Dynamics. A single MNase-seq experiment can be analysed, obtaining: nucleosome calls with nucleR, their fuzzy/well-positioned classification and stiffness estimation, Nucleosome Free Regions location, classification of TSS according to -1 and $+1$ nucleosomes, and nucleosome phasing along the gene body. Comparing two MNase-seq experiments, NucDyn identifies hotspots of changes (SHIFT $+$, SHIFT $-$, INCLUSION and EVICTION), and reports a significance score of the difference in the coverage profiles at base-pair level. Summary statistics per gene as well as genome-wide are also reported for each calculation.

Table 1. Implementation models for Nucleosome Dynamics

Code distribution		
Standalone installation	Nucleosome Dynamics CLI	https://github.com/nucleosome-dynamics/nucleosome_dynamics
	nucleR R package	https://github.com/nucleosome-dynamics/nucleR Bioconductor:
Containerized installation	NucDyn R package	http://bioconductor.org/packages/nucleR/
	Docker	https://github.com/nucleosome-dynamics/NucDyn Bioconductor: (in review)
	Singularity	https://github.com/nucleosome-dynamics/docker Docker-hub: mmbirb/nucleosome-dynamics
Platforms in use		
MuG Virtual Research Environment		https://vre.multiscalegenomics.eu/workspace/?from=nucdynwf
		https://dev.usegalaxy.es (in development) Galaxy ToolShed:
		https://toolshed.g2.bx.psu.edu/repository?repository_id=822e9e879cf92fd0

A ‘phased’ gene is defined when the distance between $+1$ and $-last$ nucleosome is close to a multiple of T (Supplementary Figure S2B). An ‘antiphased’ gene is defined when the modulus of the ratio between the distance of the $+1$ and $-last$ nucleosomes and T is close to $T/2$ (Supplementary Figure S2C). The package provides a theoretical nucleosome map based on signal propagation theory with $+1$ and $-last$ nucleosomes as emitting sites. Comparison of the predicted and the real nucleosome maps helps to detect anomalies in

the gene nucleosome distribution emerging from interacting proteins or from the effect of the remodelling machinery.

Nucleosome stiffness. Nucleosome Dynamics also analyses the sliding propensities of nucleosomes by computing its apparent resistance to be displaced along the sequence. For this purpose, we map the original reads around located nucleosomes and estimate the normalized Gaussian that better fits the distribution of reads (see Supplementary Figure

4 Nucleic Acids Research, 2019

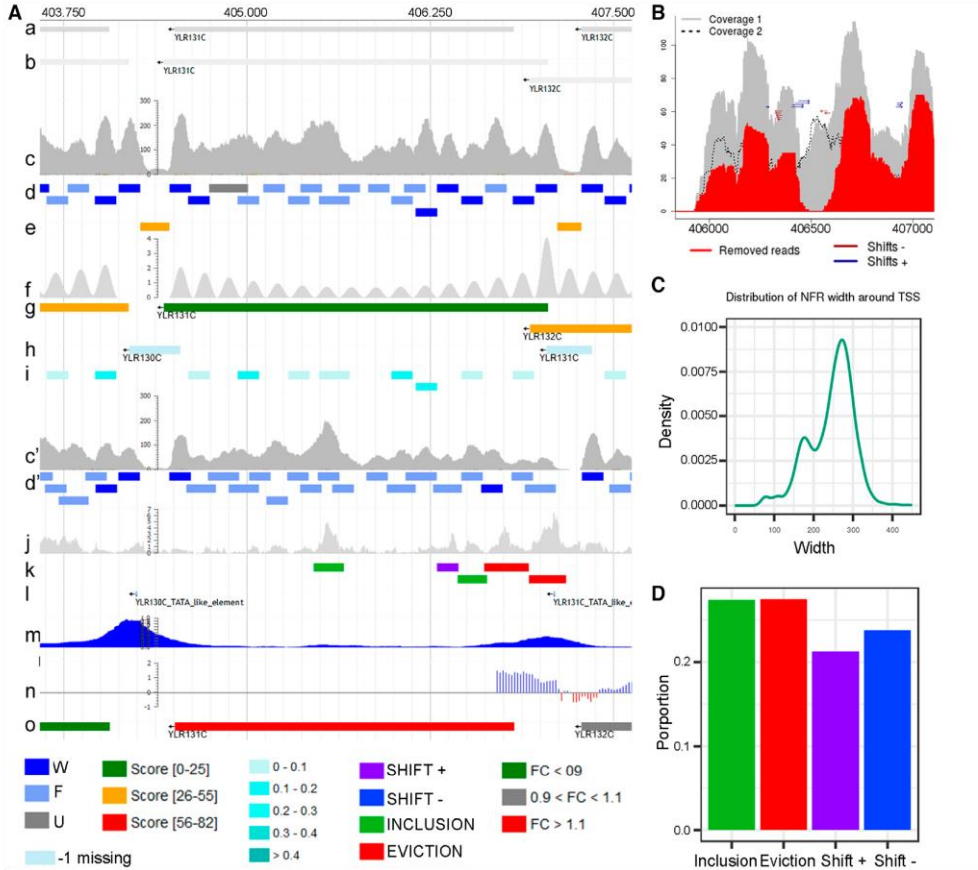


Figure 2. Visualization of Nucleosome Dynamics results in MuGVRE. (A) Nucleosome positioning along ACE2 (YLR131C) gene from *S. cerevisiae* between G2 and M cell cycle phases (a) YLR131C open reading frame; (b) ACE2 full length transcripts; (c, c') coverage of MNase-seq reads aligned to reference genome (in G2 and M phase, respectively); (d, d') nucleosome calls obtained with nucleR (G2 and M phase, respectively); (e) NFR coordinates in G2 phase; (f) prediction of the nucleosome coverage along each gene, using signals from +1 and -last nucleosomes; (g) genes shown as coloured boxes according to the phasing between the +1 and -last nucleosomes; (h) TSS classification based on nucleosomes -1 and +1 (W, F, missing) and the distance between them (open or close) represented as coloured boxes, with an arrow indicating the direction of the gene; (i) nucleosomes are coloured by their apparent stiffness value: darker blue nucleosomes are more stiff and lighter are less stiff; (j) significance of the differences in nucleosome coverage between G2 and M phases ($-\log_{10}$ of the p-value) (k) movement hotspots represented as colour coded boxes: purple for shift +, blue for shift -, green for inclusion and red for eviction; (l, m, n, o) Tracks from publicly available data representing (l) TATA elements (Rhee *et al.*, 2012), (m) TFIIB binding sites (Mayer *et al.*, 2010), (n) H3K4me3 histone mark enrichment (Liu *et al.*, 2005) and (o) gene expression changes during cell cycle (Deniz *et al.*, 2016). (B) Detailed view around a hotspot identified by NucDyn in chrXII. (C) Genome wide statistics of NFR width around TSS, in G2 phase. (D) Genome wide frequency of changes detected by NucDyn between G2 and M.

S2) from which stiffness is derived by the elastic approximation as shown in Equation (2):

$$\theta = 2 \frac{k_B T}{sd^2} \quad (2)$$

where $k_B T$ is the thermal energy at room temperature and sd is the standard deviation of the Gaussian fitted to reads associated to the nucleosome.

Defining changes in nucleosome distribution

Pairs of BAM files are processed to determine changes in the nucleosome architecture between two experiments. For this purpose (see Supplementary Figure S3) the program pairs the reads obtained from one experiment to the other to discard those that are unchanged. It also removes reads that share the same starting or ending point or those that can be fitted in longer read in the paired experiment, as they are likely to be generated by spurious differences in

nuclease degradation activity. The remaining reads are then paired between the two experiments using a dynamic programming algorithm designed to maximize: (i) the number of matches, (ii) the proximity in the middle points of the paired reads, (iii) the assignment of the paired reads to the same nucleosome. To achieve these objectives the dynamic programming highly penalizes gaps and scores read pairs inversely proportional to their distance, with a -infinite score when the distance between the middle point of the reads is longer than half the length of the nucleosome. The final output of the procedure is a set of read pairs shifted in one experiment with respect to the other. These shifts are accumulated to define hotspots that are further analysed to determine their statistical significance as markers of shifts in the nucleosome architecture.

A second type of changes detected by the program is related to differences in occupancy (insertions and evictions) between the two experiments, that are determined directly from the coverage. To reduce the impact of experimental noise we analyse the coverage data by computing a Z-score for every position x across the genome, normalizing it in 10 000 bp windows, which allows us to find locally normalized differences in coverage (Equation 3).

$$Z = \frac{m - E(m)}{(V(m))^{\frac{1}{2}}} \tag{3}$$

where m is the number of reads covering position x in experiment 1, $E(m) = nf$ (with f being the fraction of total reads in the window (N) that corresponds to experiment 1 (M) and n is the number of reads covering position x in both experiments) and $V(m)$ is the expected variance of a hypergeometric distribution, i.e. $V(m) = nf(1 - f)\frac{N-n}{n-1}$. Positive Z-score peaks mean that at that point the read coverage found at experiment 1 is higher than the coverage at experiment 2 and an eviction hotspot is annotated. Similarly, negative Z- peaks signal inclusions.

The statistical significance of the detected hotspots (shifts, inclusions and evictions) is scored using the P -values derived from Fisher's test from a contingency table between the reads in each experiment (columns) and the reads at a given position compared to reads within the window (rows):

	Exp 1.	Exp. 2	Total
Covering x	$M - m$	$N - M - n + m$	$N - n$
Not covering x	m	$n - m$	n
Total	M	$N - M$	N

Software availability and implementation

The Nucleosome Dynamics package is available in different deployment models to fulfil the needs of different users. Moreover, it is also offered as a service in two different research platforms. All available distributions are explained at Nucleosome Dynamics landing page: <http://mmb.irbbarcelona.org/NucleosomeDynamics/>, and summarized in Table 1.

Code distributions.

- Nucleosome Dynamics is written in R and composed of two packages (nucleR and NucDyn), and a series of R wrappers providing a unified interface to such core functionalities and other additional analyses (TSS classification, NFR, Phasing, Stiffness, etc., see above). Source code and documentation are available for standalone installation (see Table 1). Both nucleR and NucDyn packages are also distributed via BioConductor. Although the native R interface is recommended for experienced R users, other deployments built on top of the R software are also provided for further accessibility and portability.
- Nucleosome Dynamics package depends on a series of other R packages and helper applications. To minimize the possibilities of collision with existing installations, and to avoid installation issues to the non-experts, the packages are offered as software containers in both, the well-known Docker implementation and the Singularity format, the latter intended for multi-user systems where running Docker containers natively is not trivial – i.e. HPC systems. A single container allows the user to obtain all functions of the package directly from the command-line, and additionally, the launcher is able to accept a list of Nucleosome Dynamics analysis commands in bash to orchestrate a custom workflow. Furthermore, the use of the containers allows seamless software update. The images are registered at the corresponding hubs (see Table 1).

Use in research platforms.

- MuG Virtual Research Environment (MuGVRE) is an integrated workspace designed to put together a series of applications related to the study of 3D/4D genomics (30). The MuGVRE workspace allows to combine data, either uploaded to the workspace or obtained from public repositories such as ArrayExpress (33). MuGVRE includes applications covering a wide range of levels in the study of chromatin, from atomistic simulation or protein-nucleic acids docking to coarse-grained simulation of large nucleic acids molecules or chromatin fibers, as well as the analysis of Hi-C data. All those applications share a common data space where interoperability is assured through a common data model, and a specific protocol to incorporate new tools. MuGVRE is a cloud-based application that simplifies the deployment and provides user access to visualization tools, additional data in external repositories, and to a variety of other programs for the analysis of chromatin at different levels of resolution.
- The server provides a graphical interface based on an embedded sequence browser, Jbrowse (34), that allows visualization of nucleosome architectures in the context of other omics data (see Figure 2A). For this purpose, the outputs of all calculations are generated in GFF3 or bigWig format. Nucleosomes are represented as boxes coloured with different tones of blue according to their positioning score, with regions where nucleosome are too fuzzy displayed in a lighter colour. Similarly, nucleosome-

6 Nucleic Acids Research, 2019

free regions are highlighted by yellow boxes in a different row (see Figure 2A). In both cases, numerical information (scores, characteristics of the nucleosome or NFR) can be obtained by clicking on the corresponding boxes. The nucleosome architecture around TSS is classified based on the length of the NFR and the location score of the +1 and -1 nucleosome (see above). The results are shown as boxes between the -1 and +1 nucleosome dyads, color-coded by the corresponding architecture class. The analysis of nucleosome phasing generates a bigWig file with the theoretical prediction of the nucleosome positions inside the gene body, based on periodicity considerations and the +1 and -last nucleosomes (see above), and a GFF3 file which is displayed as a coloured box indicating whether the gene shows 'phased', 'antiphased' or intermediate nucleosome phasing (see Figure 2A). Similarly, the stiffness associated to the nucleosomes is represented (through a GFF3 file) as a box mapping to the nucleosome, coloured according to the estimated stiffness (see Figure 2A). Nucleosome Dynamics data can be put into genomic context by a series of additional tracks (Supplementary Table S3) providing gene annotations and relevant literature data. Further analyses can be obtained from the web server, such as detailed plots of nucleosome coverage and changes in nucleosome distribution between the two experiments (Figure 2B), genome-wide statistics of nucleosome architecture around TSS (Figure 2C), and overall frequency of inclusions, evictions and shifts between the two experiments (Figure 2D). Finally, Nucleosome Dynamics also generates a table listing the number of nucleosomes, their status (fuzzy/well-positioned), the identified nucleosome changes (inclusions, evictions, shifts), the classification of the promoter and the width of the NFR at the TSS for every single gene in the genome (Supplementary Table S4). These analyses are useful to test the effect of a treatment/growth conditions on nucleosome positioning both globally, or at gene level.

- **Running Nucleosome Dynamics on the galaxy platform:** Galaxy is a web-based scientific analysis platform widely used by scientists to analyse biomedical datasets such as genomics, proteomics, metabolomics or imaging (31). Nucleosome Dynamics docker has been wrapped in a series of Galaxy tools, one for each analysis. Users can launch them individually, or as part of a Galaxy workflow, building a custom pipeline that may integrate other Galaxy applications. The tools are published in the Galaxy ToolShed (see Table 1) and adopted by the ELIXIR-ES Galaxy server (currently in development phase), together with a complete ready-to-use Nucleosome Dynamics workflow. The output calculations, mainly GFF3 and bigWig files, are treated in the platform as any other sequence annotation file. Plain files like GFFs are locally displayed using a column-based visualization, while the genomic-context analysis is based on the UCSC genome browser (35). Galaxy transparently loads the data to the central UCSC browser service, and there, the sequences are loaded as custom tracks and visualized together with the other UCSC annotations.

Benchmarking data sets

The ability of the package to determine the location and nature of nucleosomes and nucleosome architectures was evaluated using synthetic maps from single cell nucleosome architectures that are combined to create *in silico* MNase digestion maps approaching closely to those found in real yeast MNase-seq experiments. For this purpose, we created multiple single cell nucleosome architectures by first placing NFRs at specific positions separated by ~2000 bp (the typical range of NFR-NFR distances in yeast) in a 10 kb DNA fragment. As the NFR are highly conserved, their positions are located with small noise in the different cells. Once the NFRs for a single cell have been placed, we defined windows for nucleosome positioning using the known average nucleosome periodicity (165 bp for yeast in the experiments simulated here). Each window was associated to either a W or F nucleosome following probability functions reproducing their expected populations at different distances from NFR. Windows that (in a given cell) appeared to be associated to W nucleosomes have a high probability to be occupied by a nucleosome, which is placed within a narrow range from the centre of the window (see Supplementary Figure S4). On the contrary, windows associated to F nucleosomes have a higher probability to be empty in a given cell, and once the nucleosome is assigned, its position is variable within the window. Once obtained, the population of *in silico* cells was processed by *in silico* MNase digestion repeating this process many times, introducing 'digestion noise' to reproduce the distributions of read lengths observed in typical yeast experiments. The integration of the reads for the entire population provides an *in silico* MNase-seq map where we know exactly the real population and fuzziness of all nucleosomes at all positions in the pool of cells. This constitutes an unambiguous benchmark to validate the performance of nucleosome annotation software. The probability functions used to generate the different cell nucleosome architectures were adjusted to qualitatively reproduce reads obtained in real MNase-seq experiments (Supplementary Figure S4). Synthetic data simulating ATAC-seq experiments can be derived in a very similar manner.

The synthetic data obtained as explained above were the starting point to generate pairs of *in silico* experiments simulating changes in nucleosome architectures. To this end, a percentage of the reads was either shifted or removed for a given nucleosome. Shifts from 1 to 5 turns of DNA (1 DNA turn = 10 bp) were introduced generating 100 replicates in each case to evaluate the sensitivity of the method to detect shifts of different lengths and affecting different percentage of the total population.

RESULTS**Performance using *in silico* datasets**

NucleR. We explored the ability of the software to position nucleosomes using as reference our highly controlled synthetic data (see Methods). As a control, we repeated the exercise using another widely used program for nucle-

osome annotation, DANPOS (15). Both packages show a good ability to represent the nucleosome architecture using MNase-seq data. In terms of occupancy DANPOS performs slightly better than the nucleR module implemented in our Nucleosome Dynamics package ($R^2 = 0.97$ for DANPOS versus $R^2 = 0.93$ for nucleR), while nucleR can detect better the nucleosome fuzziness ($R^2 = 0.94$ for nucleR versus $R^2 = 0.87$ for DANPOS; see Supplementary Methods, for description of the metrics). The location of W nucleosomes is nearly identical in both methods, but for F nucleosomes the results are quite different, as DANPOS annotates a wide region of sequence reads as a single nucleosome positioned with a large uncertainty, while nucleR can assign several nucleosomes to the wide signal, even when in some cases the two nucleosomes can partially overlap (see Figure 3A). As a result, DANPOS provides, probably, the best 'average' distribution of nucleosomes, but nucleR provides a more realistic picture of the cellular variability, capturing the presence of alternative nucleosome architectures in the cellular population. As it can be seen in Figure 3A, where selected examples of DANPOS and nucleR nucleosome distributions are compared with the real nucleosome architecture existing in our synthetic data; in Figure 3B, where we compare the ability of DANPOS and nucleR to detect the presence of a percentage of cells showing a different nucleosome architecture and in Figure 3C, where we report the average distance between the real position of the dyads of the synthetic nucleosomes and those located by DANPOS or nucleR.

NucDyn. We tested the ability of our method to detect rearrangements in the nucleosome architecture using again our controlled *in-silico* MNase-seq data, simulating displacement (shift), insertion or eviction of one nucleosome, occurring in a different percentage of the cells. Sizeable changes such as nucleosome insertion or eviction are detected with good sensitivity by our method, even when they affect a relatively small percentage of cells (Figure 3D), while DANPOS or Dimnp only detect such changes when affecting a very large proportion of cells. Small nucleosome shifts (implying less than one turn of DNA) are not detectable by our algorithm unless they occur in a large percentage of the cells; while shifts implying a displacement of at least two turns of DNA (20 base pairs) are detectable with good sensitivity, even when affecting less than half of the cellular population (see Figure 3E). In this case, the comparison with other programs is difficult, as only DANPOS (15) allows an indirect way to detect nucleosome shifts by looking at distances between nucleosome peaks in both experiments. Unfortunately, with our synthetic data, DANPOS achieved poor sensitivity (less than 0.20 for 5 turns of DNA shift in 70% of cells and <0.1 for 3 turns shifts affecting also 70% of cells; see Supplementary Figure S5).

In summary, analysis of well-controlled *in silico* data shows that the Nucleosome Dynamics package (including NucDyn and nucleR) is not only a very powerful tool to define nucleosome families from MNase-seq experiments performed with a population of cells, but also a powerful approach to detect subtle changes in nucleosome architecture affecting a percentage of the cells in the studied sample.

Test cases

In order to illustrate the information derived from Nucleosome Dynamics, we applied our method in different real cases where experimental MNase-seq data were available. It is important to mention that the biological relevance of this type of comparison depends on the quality of the data and especially on the similar level of MNase digestion of the samples being compared. Indeed, several groups, including ours, have demonstrated the impact of the level of MNase digestion on the final nucleosome maps in several organisms, essentially at the level of the so-called 'fragile' nucleosomes (19,36–38). To illustrate this observation, we took advantage of the extensive study made by (36) and used Nucleosome Dynamics to compare nucleosome positioning in the input of two H2B and two H4 MNase-ChIP-seq samples, one under-digested and one over-digested (50U and 400 U of MNase respectively for H2B; 25 U and 300 U MNase respectively for H4). First, we focused on the H2B-input samples and confirmed that the number of nucleosome detected by nucleR decreases as the amount of MNase increases (from 80 160 down to 72 775, Supplementary Table S5) which is corroborated by the detection by NucDyn of 3559 evictions genome wide (Supplementary Table S6). At the promoter level, the proportion of W-open-W TSS increases from 123 to 2026 while the W-close-W TSS decreases from 2656 down to 346 (Supplementary Figure S6A). Regarding the phasing analysis, the percentage of phased genes does not change significantly due to the level of MNase digestion (Supplementary Figure S6B). Similar numbers were obtained for the H4-input samples. Hence, it is important to control MNase digestion level before using Nucleosome Dynamics package. Another technique that is not influenced by the level of MNase digestion is chemical cleavage mapping. NucleR can be applied to map nucleosome positions using the coverage obtained from these experiments (Supplementary Figure S7).

Cell cycle. The first example comes from the analysis of the changes in nucleosome organization occurring along the cell cycle in yeast, using our own previously published data (39). As shown in Figure 4A, the nucleR module of the Nucleosome Dynamics package suggests that nucleosomes tend to be fuzzier (F) in S and M phases compared to G1 and G2 phases. The increase in fuzziness in S and M phases impacts on the promoter classification as the number of W-open-W and W-closed-W promoters decreases compared to G1 and G2 (Figure 4B), but overall the ratio of closed/open NFRs (nucleosome free regions) is not dramatically altered along cell cycle (Supplementary Figure S8). Very interestingly, the changes in nucleosome architecture detected by NucDyn are not randomly distributed along the genome, but appear to be localized in specific families of genes, which are related to cell cycle progression, as shown by Gene Ontology (GO) Enrichment Analysis (40) in Figure 4C. Examples of the detailed information provided by Nucleosome Dynamics for some specific genes are shown in Figure 4D, where we report nucleosome maps of *PRY2* (a gene related to lipid transport), whose expression peaks in G1 phase, *YHP1* (involved in negative regulation of transcription of certain cell cycle genes), and *GIC1* (a GTPase-interaction

8 Nucleic Acids Research, 2019

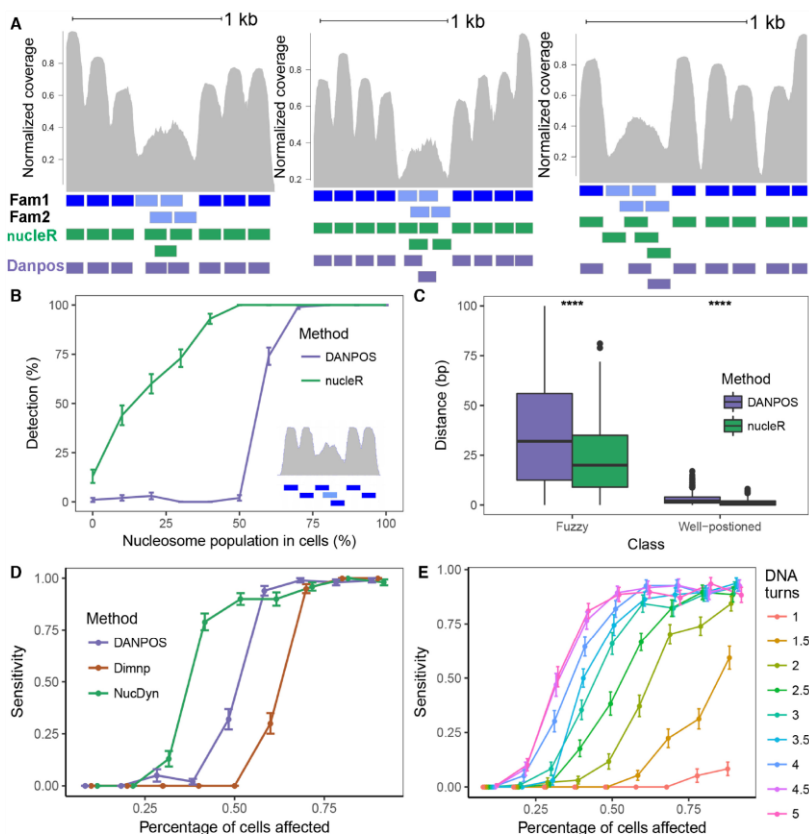


Figure 3. Performance of nucleR and NucDyn. (A) Coverage of three synthetic nucleosome maps (shown in grey), containing well-positioned (dark blue) and fuzzy nucleosomes (light blue). Two possible nucleosome families generate different nucleosome positioning in fuzzy regions (Fam1 and Fam2). Predicted nucleosome positions using nucleR and Danpos are shown in green and purple, respectively. (B) Comparison of nucleR and Danpos for detection of a second family of nucleosomes (light blue nucleosome in the bottom-right panel). Y-axis shows the number of cells required in the second family in order to be detected by the algorithm. (C) Distance between the dyads identified by nucleR (green) and DANPOS (purple) to the dyad position in the true synthetic nucleosome map for fuzzy and well positioned nucleosomes. (D) Sensitivity of the EVICTION prediction for NucDyn, DANPOS and Dimnp. Evictions were simulated removing reads from a given percentage of families (10%, 20%, ..., 90%) and were identified from DANPOS output as a nucleosome with $\text{point_log}_2\text{FC} < -1$ and $\text{point_diff_FDR} < 0.01$ (point with highest difference in the two samples, as reported by the software), and with default parameters for Dimnp. (E) Sensitivity of the SHIFT prediction computed on synthetic nucleosome maps. Shifts were introduced displacing reads from 1 to 5 DNA turns and modifying different percentages of the families (10%, 20%, ..., 90%).

component involved in mitosis regulation) expressed in S phase (39). In the three cases, expression changes correlate with significant variation in nucleosome architecture at the promoter region between two stages of cell cycle. Typically, eviction or shifts reducing the presence of nucleosome in the core promoter region are related to active states of the genes (4,19,28,41,42).

Yeast metabolic cycle. A second example of use of our tool was the comparison of nucleosome architecture amongst cells at different stages of the yeast metabolic cycle (YMC). We took advantage here of high resolution experiments (43) in which the authors analysed simultaneously gene expres-

sion and MNase-seq maps at regularly spaced periods of time after adding fresh culture media. At two of these time points (T9 and T12 in Nocetti and Whitehouse 2016) dramatic changes of expression in genes related to reductive charging (poorly transcribed at T9 and highly transcribed at T12) and oxidative phase (highly transcribed at T9 and poorly transcribed at T12) have been detected. Analysis of global nucleosome architecture shows moderate changes between T9 and T12 (Figure 5A), but differences are more noticeable when the analysis is focused on Ox-genes (involved in amino acid synthesis, sulphur metabolism, ribosome and RNA metabolism (44), which are expressed in T9 and repressed in T12) and R/C genes (involved in non-

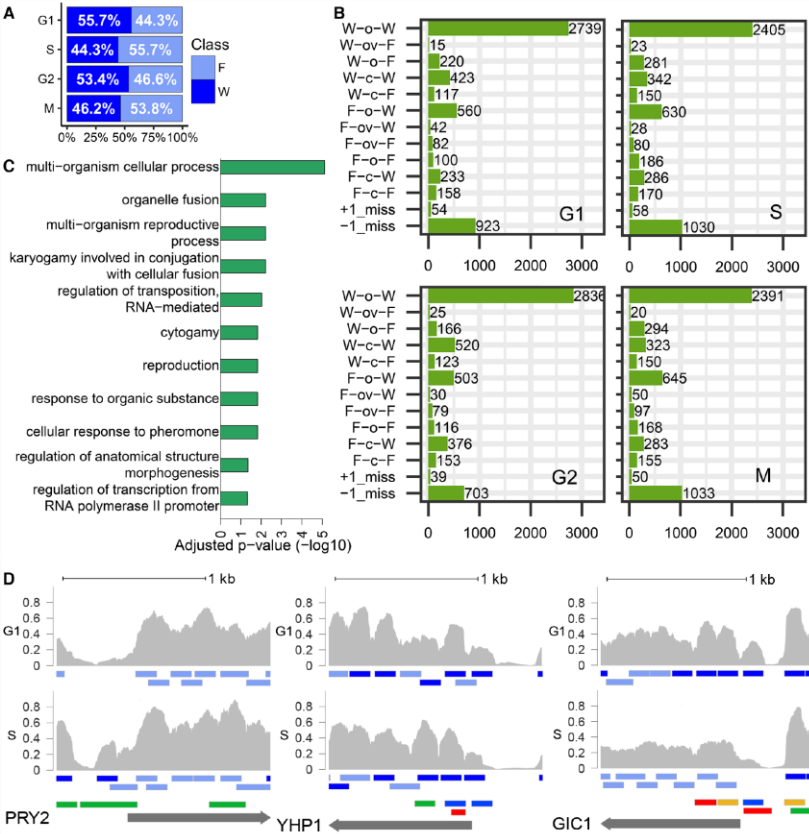


Figure 4. Nucleosome Dynamics along the cell cycle. (A) Percentage of fuzzy and well-positioned nucleosomes and (B) promoter classification (number of genes in each class) for every cell cycle stage. (C) GO terms enriched in genes with nucleosome changes between G1 and S detected by NucDyn. (D) Example of three cell-cycle dependent genes that present differential nucleosome architectures between G1 and S. In gray, the normalized coverage from the BAM files of the two cell cycle stages, 500 bp upstream and 1000 bp downstream the TSS. Below each BAM file, the nucleosome calls obtained with nucleR are represented (dark blue for well-positioned nucleosomes, light blue for fuzzy nucleosomes). The fifth track contains shifts (yellow for positive, blue for negative), inclusions (green) and evictions (red) identified by NucDyn.

respiratory metabolism, protein degradation, autophagy and vacuole (44), which are expressed in T12 and repressed in T9). Nucleosome Dynamics allows the detection and quantification of the alterations in nucleosome architecture coupled to such changes in expression. Thus, for Ox-genes (Figure 5B), the fuzziness at the -1 nucleosome decreases when moving from T9 to T12, in agreement with the general rule that reduced NFR upstream the well positioned $+1$ nucleosome correlates with inactive genes. On the contrary, for R/C genes (Figure 5C) the NFR upstream well positioned $+1$ nucleosome enlarges, since the -1 nucleosomes become fuzzier, again in perfect agreement with the changes of expression. To discard any biases resulting from the MNase digestion conditions, we confirmed that the length of the sequenced fragments was comparable in both samples (Sup-

plementary Figure S9). Examples of the detailed information provided by Nucleosome Dynamics for three Ox-genes (*NSR1*, *RRP4* and *SNU13*, all of them related to ribosomal biogenesis and RNA metabolism) are shown in Figure 5D, where upon T9→T12 transition, shifts and even insertions are shown leading to a reduction in the width of the NFR upstream the TSS: a fingerprint of gene deactivation. Similarly, Figure 5E provides the same type of information for three R/C genes (*CTAI*, *SUE1* and *SAFI*), which are associated respectively with peroxisome, cytochrome C degradation and proteasome (see above). In the three cases, T9→T12 transition is coupled with massive nucleosome eviction upstream the TSS, leading to open configurations of NFR, typical of highly expressed genes.

10 Nucleic Acids Research, 2019

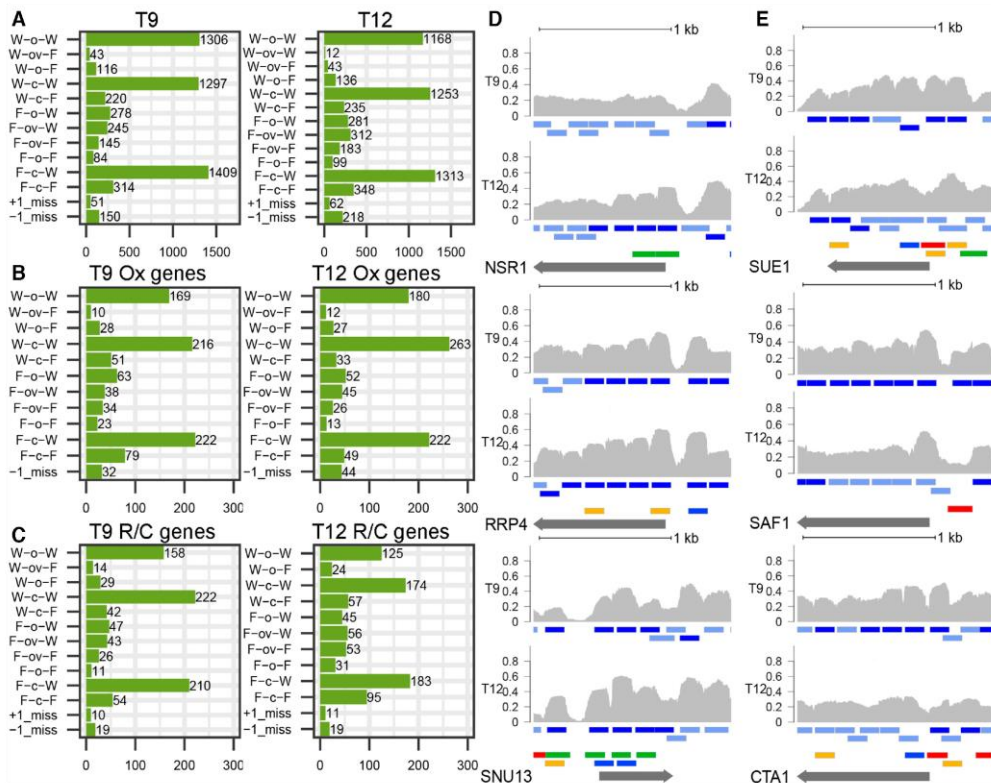


Figure 5. Nucleosome Dynamics in two points of the YMC. Promoter classification in the two time points (T9 and T12) for (A) all genes, (B) genes from the Ox cluster and (C) genes from the R/C cluster. (D and E) Example of three genes from Ox and 3 from R/C clusters that present differential nucleosome architectures between T9 and T12. In grey, the normalized coverage from the BAM files of the two time points, 500 bp upstream and 1000 bp downstream the TSS. Below each BAM file, the nucleosome calls obtained with nucleR are represented (dark blue for well-positioned nucleosomes, light blue for fuzzy nucleosomes). The fifth track contains shifts (yellow for positive, blue for negative), inclusions (green) and evictions (red) identified by NucDyn.

Changes in nutrients. As a last test, we applied Nucleosome Dynamics to explore the modifications in nucleosome architecture in yeast, linked to the change in the media from glucose-rich to either galactose-rich or ethanol-rich (45). Changes in the TSS nucleosome architecture classification occur among the three conditions (Figure 6A). There are not complete expression data in Kaplan *et al.* 2009, but we expected that replacement of glucose by galactose in the media would imply changes in expression in genes related to carbohydrate metabolism and transport which encouragingly, are those where sizeable changes in nucleosome architecture are detected by Nucleosome Dynamics (Figure 6B). Similarly, replacement of glucose by ethanol was expected to have an impact on the cell through: (i) expression of stress response genes, (ii) changing completely hexose metabolism in the absence of hexoses and (iii) eliminating ethanol through oxidation generating changes in the redox state of the cell which need to be corrected (46). Very encouragingly again, genes involved in stress response, hex-

ose metabolism and redox activities are those for which the largest changes in nucleosome architecture have been detected (Figure 6C).

We analysed in detail some genes which are expected to change dramatically their expression upon glucose→galactose substitution, as they are crucial to integrate galactose in normal hexose metabolism: *GAL1*, coding for a Galactokinase, *GAL10*, coding for the UDP-glucose-4-epimerase, and *GAL7*, coding for the galactose-1-phosphate uridyl transferase (Figure 6D). In the three cases evictions and shifts (in some cases noticeably) generate wider NFRs upstream the gene, changes that in some cases extend to the coding regions and that signal a pronounced increase in expression of these galactose-related genes.

A similar detailed analysis was made for three genes which are expected to be overexpressed when ethanol substitutes glucose as energy source: two stress response genes *HSP26* and *HSP12*, and *HXK1*, a hexokinase activated when cells are shifted to a non-fermentable carbon source

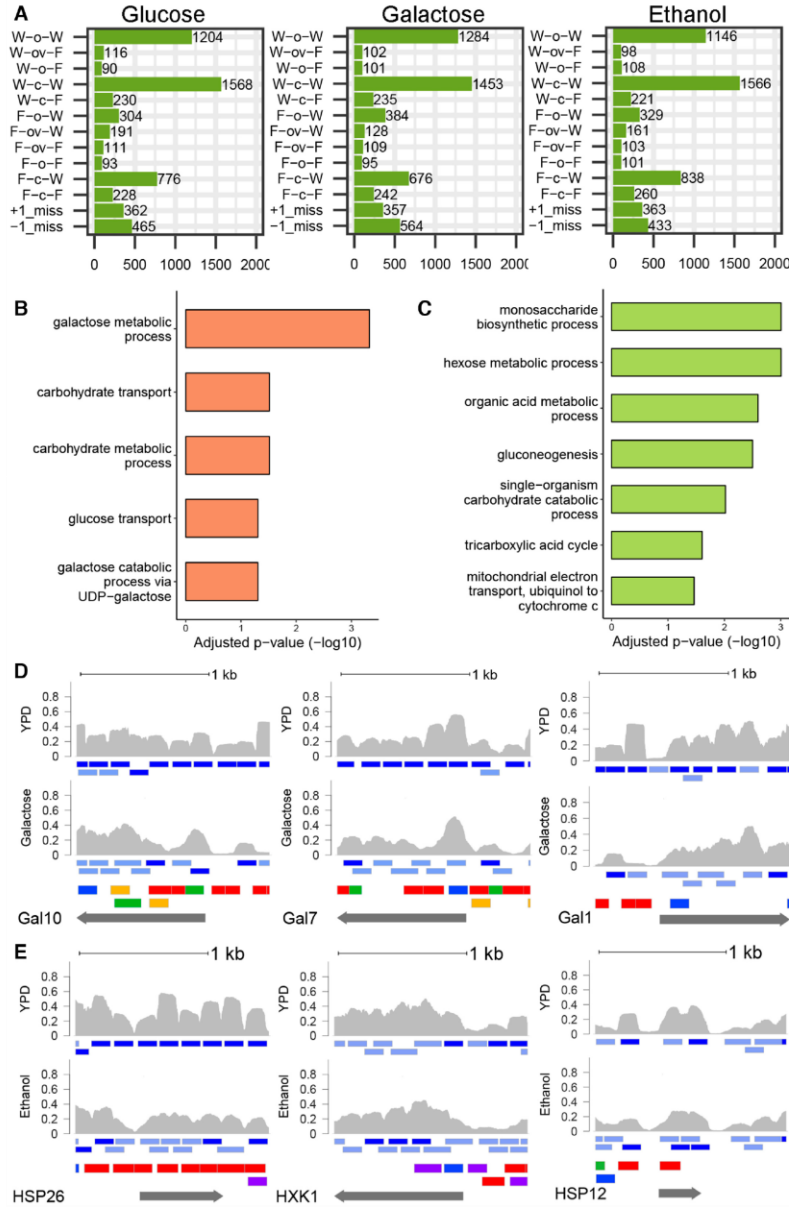


Figure 6. Nucleosome Dynamics under different nutrient conditions. (A) Promoter classification in glucose, galactose and ethanol rich media. (B and C) GO terms enriched in genes with nucleosome changes detected by NucDyn, changing the medium from glucose to galactose or ethanol, respectively. (D and E) Example of three genes involved in galactose and ethanol metabolism, respectively, that present differential nucleosome architectures depending on the carbon source. In gray, the normalized coverage from the BAM files of the two cell cycle stages, 500 bp upstream and 1000 bp downstream the TSS. Below each BAM file, the nucleosome calls obtained with nucleR are represented (dark blue for well-positioned nucleosomes, light blue for fuzzy nucleosomes). The fifth track contains shifts (yellow for positive, blue for negative), inclusions (green) and evictions (red) identified by NucDyn.

12 *Nucleic Acids Research*, 2019

such as ethanol (47). Results in Figure 6E illustrate the magnitude of the changes (mainly evictions) detected by Nucleosome Dynamics, which affect the NFR, and even in some cases the coding regions.

DISCUSSION

Different studies demonstrated that nucleosome architecture is coupled to gene function (4,43,48) and that transcriptional activity and nucleosome architecture are tightly coupled. Unfortunately, detecting changes in nucleosome architecture is complex as nucleosomes are dynamic and even a population of 'identical' well synchronized and grown under identical conditions cells might have nucleosomes placed at different positions. This, combined with the intrinsic uncertainties of MNase- or ATAC-seq experiments, generate noisy data which are difficult to process for precisely locating nucleosomes and even more difficult to detect significant changes in nucleosome arrangements due to internal or external stimuli. The suite of programs incorporated in Nucleosome Dynamics allows not only a robust location of nucleosomes, even in cases of heterogeneous pools of cells, but also the detection of changes in nucleosome arrangements, even those affecting a moderate population of cells. To increase its utility, Nucleosome Dynamics is integrated into a powerful virtual research environment, where it is combined with different tools for analysis of data and visualization in the context of genomic metadata, which help the user not only to analyse nucleosome architecture and dynamics, but also to put them in the context of known genomic information (Figure 2).

We validated the methodology using synthetic data that mimic typical MNase-seq maps, in which the positions of the nucleosome in the different cells are unambiguously known. The two main modules of Nucleosome Dynamics (nucleR and NucDyn) perform very well capturing cellular diversity and detecting shifts, evictions and inclusions that affect a moderate percentage of the cellular population. Furthermore, we tested the power of the methodology by exploring nucleosome rearrangements occurring along cell cycle, yeast metabolic cycle, and those linked to the change in the source of energy from glucose to galactose or ethanol. In the tested cases, Nucleosome Dynamics provides accurate global and local descriptions of nucleosome structure and dynamics and deciphers the nature of the connection between nucleosome organization and gene expression.

DATA AVAILABILITY

Raw MNase-seq datasets reported as test cases in this study were obtained from the *ENA-SRA* website (<http://www.ebi.ac.uk/ena>) and the GEO repository under accession numbers: PRJEB6970 for the cell cycle data, GSE77631 for the yeast metabolic cycle, GSE13622 for the nucleosome maps from yeast cultivated in glucose, galactose and ethanol media, and GSE83123 for the different levels of MNase digestion. Processed chemical cleavage data was obtained from GSE97290.

The processed test data and benchmarking synthetic data supporting the conclusions of this article are available in Zenodo repository (10.5281/zenodo.2632999), and can be

incorporated to both MuGVRE and Galaxy Nucleosome Dynamics installations for additional testing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted all members of the MuG consortia for help and for acting as a beta-tester of the software.

FUNDING

M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avancats) academia researcher; Spanish Ministry of Science [RTI2018-096704-B-I00]; Catalan Government [2017-SGR-134]; Instituto de Salud Carlos III-Instituto Nacional de Bioinformática, the European Union's Horizon 2020 research and innovation program, and the Biomolecular and Bioinformatics Resources Platform [ISCH PT 17/0009/0007 co-funded by the Fondo Europeo de Desarrollo Regional FEDER; Grants Elixir-Excelerate: 676559 and BioExcel2: 823830; ERC:812850; MuG-676566]; MINECO Severo Ochoa Award of Excellence from the Government of Spain (awarded to IRB Barcelona). Funding for open access charge: Spanish Ministry of Science [RTI2018-096704-B-I00].

Conflict of interest statement. None declared.

REFERENCES

- Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-Resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
- Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venter, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C. *et al.* (2008) Nucleosome organization in the *Drosophila* genome. *Nature*, **453**, 358–362.
- Yuan, G.-C. (2005) Genome-Scale Identification of Nucleosome Positions in *S. cerevisiae*. *Science*, **309**, 626–630.
- Lai, W.K.M. and Pugh, B.F. (2017) Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat. Rev. Mol. Cell Biol.*, **18**, 548–562.
- Rando, O.J. and Ahmad, K. (2007) Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.*, **19**, 250–256.
- Jiang, C. and Pugh, B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.*, **10**, 161–172.
- Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A. and Segal, E. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.*, **44**, 743–750.
- Collings, C.K. and Anderson, J.N. (2017) Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenet. Chromatin*, **10**, 18.
- Kubik, S., O'Duibhir, E., de Jonge, W.J., Mattarocci, S., Albert, B., Falcone, J.-L., Bruzzone, M.J., Holstege, F.C.P. and Shore, D. (2018) Sequence-directed action of RSC remodeler and general regulatory factors modulates +1 nucleosome position to facilitate transcription. *Mol. Cell*, **71**, 89–102.
- Mellor, J. and Morillon, A. (2004) ISWI complexes in *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta (BBA) - Gene Struct. Expression*, **1677**, 100–112.
- Knight, B., Kubik, S., Ghosh, B., Bruzzone, M.J., Geertz, M., Martin, V., Dénervaud, N., Jacquet, P., Ozkan, B., Rougemont, J. *et al.*

- (2014) Two distinct promoter architectures centered on dynamic nucleosomes control ribosomal protein gene transcription. *Genes Dev.*, **28**, 1695–1709.
12. Whitehouse, I., Flaus, A., Cairns, B.R., White, M.F., Workman, J.L. and Owen-Hughes, T. (1999) Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature*, **400**, 784–787.
 13. Whitehouse, I., Stockdale, C., Flaus, A., Szczelkun, M.D. and Owen-Hughes, T. (2003) Evidence for DNA translocation by the ISWI chromatin-remodeling enzyme. *Mol. Cell Biol.*, **23**, 1935–1945.
 14. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
 15. Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X. and Li, W. (2013) DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.*, **23**, 341–351.
 16. Chen, W., Liu, Y., Zhu, S., Green, C.D., Wei, G. and Han, J.-D.J. (2014) Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat. Commun.*, **5**, 4909.
 17. Flores, O. and Orozco, M. (2011) nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, **27**, 2149–2150.
 18. Teif, V.B. (2016) Nucleosome positioning: resources and tools online. *Brief. Bioinform.*, **17**, 745–757.
 19. Flores, O., Deniz, O., Soler-López, M. and Orozco, M. (2014) Fuzziness and noise in nucleosomal architecture. *Nucleic Acids Res.*, **42**, 4934–4946.
 20. Deniz, Ö., Flores, O., Battistini, F., Pérez, A., Soler-López, M. and Orozco, M. (2011) Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC Genomics*, **12**, 489.
 21. Gutiérrez, G., Millán-Zambrano, G., Medina, D.A., Jordán-Pla, A., Pérez-Ortín, J.E., Peñate, X. and Chávez, S. (2017) Subtracting the sequence bias from partially digested MNase-seq data reveals a general contribution of TFIS to nucleosome positioning. *Epigenet. Chromatin*, **10**, 58.
 22. Brogaard, K., Xi, L., Wang, J.-P. and Widom, J. (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature*, **486**, 496–501.
 23. Chereji, R.V., Ramachandran, S., Bryson, T.D. and Henikoff, S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol.*, **19**, 19.
 24. Voong, L.N., Xi, L., Sebeson, A.C., Xiong, B., Wang, J.-P. and Wang, X. (2016) Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell*, **167**, 1555–1570.
 25. Thakur, J., Talbert, P.B. and Henikoff, S. (2015) Inner kinetochore protein interactions with regional centromeres of fission yeast. *Genetics*, **201**, 543–561.
 26. Moyle-Heyman, G., Zaichuk, T., Xi, L., Zhang, Q., Uhlenbeck, O.C., Holmgren, R., Widom, J. and Wang, J.-P. (2013) Chemical map of *Schizosaccharomyces pombe* reveals species-specific features in nucleosome positioning. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 20158–20163.
 27. Henikoff, S., Ramachandran, S., Krassovsky, K., Bryson, T.D., Codomo, C.A., Brogaard, K., Widom, J., Wang, J.-P. and Henikoff, J.G. (2014) The budding yeast Centromere DNA Element II wraps a stable Cse4 hemisome in either orientation in vivo. *eLife*, **3**, e01861.
 28. Liu, L., Xie, J., Sun, X., Luo, K., Qin, Z.S. and Liu, H. (2017) An approach of identifying differential nucleosome regions in multiple samples. *BMC Genomics*, **18**, 135.
 29. R Core Team. (2016) *R-A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
 30. Codó, L., Bayarri, G., Cid-Fuentes, J.A., Conejero, J., Hospital, Adam, Royo, R., Repchevsky, D., Pasi, M., Meletiou, A., McDowall, M.D. et al. (2019) MuGVRE. A virtual research environment for 3D/4D genomics. *Bioinformatics*.
 31. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A. et al. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
 32. Giancoli, D.C. (2000) *Physics for Scientists & Engineers with Modern Physics*. 3rd edn. Prentice Hall, Upper Saddle River.
 33. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T. et al. (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
 34. Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elisk, C.G., Lewis, S.E., Stein, L. et al. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
 35. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, A. D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
 36. Chereji, R.V., Ocampo, J. and Clark, D.J. (2017) MNase-sensitive complexes in yeast: nucleosomes and non-histone barriers. *Mol. Cell*, **65**, 565–577.
 37. Chereji, R.V., Kan, T.-W., Grudniewska, M.K., Romashchenko, A.V., Berezikov, E., Zhimulev, I.F., Guryev, V., Morozov, A.V. and Moshkin, Y.M. (2016) Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in *Drosophila melanogaster*. *Nucleic Acids Res.*, **44**, 1036–1051.
 38. Jeffers, T.E. and Lieb, J.D. (2017) Nucleosome fragility is associated with future transcriptional response to developmental cues and stress in *C.elegans*. *Genome Res.*, **27**, 75–86.
 39. Deniz, Ö., Flores, O., Aldea, M., Soler-López, M. and Orozco, M. (2016) Nucleosome architecture throughout the cell cycle. *Sci. Rep.*, **6**, 19729.
 40. Falcon, S. and Gentleman, R. (2007) Using GOSTats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
 41. Teif, V.B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.-P., Marth, C., Höfer, T. and Rippe, K. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.*, **19**, 1185–1192.
 42. Chen, J., Li, E., Zhang, X., Dong, X., Lei, L., Song, W., Zhao, H. and Lai, J. (2017) Genome-wide nucleosome occupancy and organization modulates the plasticity of gene transcriptional status in maize. *Mol. Plant*, **10**, 962–974.
 43. Nocetti, N. and Whitehouse, I. (2016) Nucleosome repositioning underlies dynamic gene expression. *Genes Dev.*, **30**, 660–672.
 44. Tu, B.P., Kudlicki, A., Rowicka, M. and McKnight, S.L. (2005) Logic of the yeast metabolic Cycle: Temporal compartmentalization of cellular processes. *Science*, **310**, 1152–1158.
 45. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
 46. Stanley, D., Bandara, A., Fraser, S., Chambers, P.J. and Stanley, G.A. (2010) The ethanol stress response and ethanol tolerance of *Saccharomyces cerevisiae*. *J. Appl. Microbiol.*, **109**, 13–24.
 47. Rodríguez, A., Cera, T. de la, Herrero, P. and Moreno, F. (2001) The hexokinase 2 protein regulates the expression of the GLK1, HXK1 and HXK2 genes of *Saccharomyces cerevisiae*. *Biochem. J.*, **355**, 625–631.
 48. Bai, L. and Morozov, A.V. (2010) Gene regulation by nucleosome positioning. *Trends Genet.*, **26**, 476–483.

Bibliography for Chapter 5

- [1] W. K. M. Lai and B. F. Pugh, "Understanding nucleosome dynamics and their links to gene expression and DNA replication," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 9, pp. 548–562, May 2017.
- [2] L. Bai and A. V. Morozov, "Gene regulation by nucleosome positioning," *Trends Genet.*, vol. 26, no. 11, pp. 476–483, Nov. 2010.
- [3] O. Flores and M. Orozco, "nucleR: a package for non-parametric nucleosome positioning," *Bioinformatics*, vol. 27, no. 15, pp. 2149–2150, Aug. 2011.
- [4] K. Chen *et al.*, "DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing," *Genome Res.*, vol. 23, no. 2, pp. 341–351, Feb. 2013.
- [5] L. Liu, J. Xie, X. Sun, K. Luo, Z. S. Qin, and H. Liu, "An approach of identifying differential nucleosome regions in multiple samples," *BMC Genomics*, vol. 18, no. 1, Dec. 2017.
- [6] L. Codó *et al.*, "MuGVRE. A virtual research environment for 3D/4D genomics," *Bioinformatics*, preprint, 2019.
- [7] E. Afgan *et al.*, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W537–W544, Jul. 2018.
- [8] K. Brogaard, L. Xi, J.-P. Wang, and J. Widom, "A map of nucleosome positions in yeast at base-pair resolution," *Nature*, vol. 486, no. 7404, pp. 496–501, Jun. 2012.
- [9] R. V. Chereji, S. Ramachandran, T. D. Bryson, and S. Henikoff, "Precise genome-wide mapping of single nucleosomes and linkers in vivo," *Genome Biol.*, vol. 19, no. 1, Dec. 2018.

Chapter 6 . Impact of DNA methylation on 3D genome structure

DNA methylation is a well-known epigenetic mark implicated in development and disease [1]. It has been shown that CpG methylation affects the physical properties of DNA, increasing its stiffness, which in turns affects nucleosome binding [2,3]. Here we present a comprehensive study about the effect of DNA methylation on chromatin structure, both at the nucleosome level and at the whole genome 3D configuration in the nucleus. Although many studies have explored the correlation between nucleosome positioning and DNA methylation, it is still unclear whether the two factors are correlated or anti correlated [2]–[4]. Moreover, although typically DNA methylation was considered a hallmark of repression at promoters [5], nowadays we know that the relationship is more complex [1,6].

The contradictory results found to this day on the relationship methylation \leftrightarrow nucleosome arrangement \leftrightarrow gene expression might be due to the presence of other methylation readers present in higher eukaryotes controlling directly or indirectly nucleosome positioning and epigenetic-dependent gene expression. For this reason, we employed a natively unmethylated genome and induced the expression of four murine DNA methyltransferases (DNMTs) to methylate the DNA. This allows to directly study the intrinsic effect of DNA methylation on the chromatin structure of this system in the absence of methylated DNA recognition proteins.

The induced methylation in yeast follows a similar pattern as in mammalian cells, lower at the TSS and increasing towards the TTS. Here, we found that in our system

all the four DNMTs are active and removing each of them, one at the time, produced a significant decrease in the level of methylation reached. In general, similar patterns of DNA methylation are observed in all cases, although we detected some specific DNA sequence preferences for DNMT3a and DNMT3b. Combining all of them we managed to obtain a very high level of DNA methylation in yeast genome. We found that nucleosome positioning strongly guides the position of the CpG methylation, since well-positioned (W) nucleosomes are depleted of methylation around the dyad accumulating towards the linkers, while fuzzy (F) nucleosomes can be methylated at equal rates at any position. On the other hand, nucleosome changes are also produced by the DNA methylation: the proportion of F nucleosomes increased, especially towards the 3' end of the genes, and at highly methylated promoters we detected differential nucleosome positioning with our Nucleosome Dynamics package (see Chapter 5 for a description of the algorithm).

Among the highly methylated promoters, we also found differential expression, which was not present in low-methylated. Some genes are repressed upon methylation, which can be explained considering the steric hindrance that a displaced nucleosome generates, but surprisingly several over-expressed genes are detected and quite interestingly are related to meiosis. We investigated the corresponding promoters finding a common motif that is CpG rich, URS1, target site for UME6 protein, known to be a repressor for meiosis-related genes. Since the level of methylation in this motif is proportional to the increase in transcription, the over expression could be explained by the unbinding of UME6 repressor. It is interesting to notice that this effect is direct, not protein-mediated and by might be related to intrinsic changes in DNA-URS1 binding related to cytosine methylation.

Next, to explore the intrinsic effect of DNA methylation on global 3D chromatin structure, we performed Hi-C experiments and developed a restraint-based chromatin model. We produced an ensemble of structures for each yeast chromosome based on restraints derived from the Hi-C contact matrices binned at 5kbp. The 3D models of several chromosomes show lower flexibility in the methylated chromatin, revealed by the lower RMSD of each structure in the ensemble and lower RMSF of each bead among the structures in the ensemble.

Globally, yeast chromatin reorganizes upon DNA methylation induction, losing interactions in *trans* and gaining contacts in *cis*, especially around the centromeres. This might be explained by the high density of chromatin in these loci, since yeast centromeres are attached to the Spindle Pole Body.

Particularly, we found large chromatin changes in chromosome XII, containing the rDNA repeats. In our unmethylated controls, cells are in stationary phase and the expression of these genes decreases, relaxing the barrier that this locus represents for separating the two sides of the chromosome, upstream and downstream from this region. In the methylated strain however, the contacts between the two parts of the chromosome are lost, suggesting that the chromatin structure is blocked upon methylation while the cells are dividing, keeping the rDNA region separated from the rest of the chromatin, as is known to occur in replicating cells [8], [9], even after the cells enter the stationary phase.

Overall, our results show the intrinsic effect of DNA methylation on structural changes in chromatin organization, independent of DNA methylation readers as our model organism does not contain the complex cellular machinery that recognizes methylation signatures.

Publication:

Diana Buitrago¹, Mireia Labrador¹, Simon Heath, Juan Pablo Arcon, Rafael Lema, Oscar Flores, Anna Esteve-Codina, Julie Blanc, David Bellido, Marta Gut, Ivo Gut, Pablo D. Dans, Isabelle Brun Heath, Modesto Orozco. Impact of DNA methylation on 3D genome structure. (*in preparation*)

Supplementary material for this article can be found in the **Annex V**.

¹ Equally contributing authors

Impact of DNA methylation on 3D genome structure

Diana Buitrago^{1,2*}, Mireia Labrador^{1,2*}, Simon C. Heath⁴, Juan Pablo Arcon^{1,2}, Rafael Lema^{1,2}, Oscar Flores^{1,2}, Anna Esteve-Codina⁴, Julie Blanc⁴, David Bellido⁵, Marta Gut⁴, Ivo Gut⁴, Pablo D. Dans^{1,2}, Isabelle Brun Heath^{1,2}, Modesto Orozco^{1,2,3}

¹ Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldori i Reixach 10. Barcelona 08028. Spain, ²Joint IRB-BSC Program in Computational Biology, Barcelona, Spain, ³Departament de Bioquímica i Biologia Molecular, Universitat de Barcelona, Barcelona, Spain, ⁴ Centro Nacional de Análisis Genómico (CNAG-CRG), Centre de Regulació Genómica (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain, ⁵Centres Científics i Tecnològics, Universitat de Barcelona, Barcelona, Spain

* These authors contributed equally to this work

Abstract

The extreme complexity of epigenetic regulation in higher organisms makes the determination of a causal function of DNA methylation complicated. We investigated the role of DNA methylation in a simpler model system, budding yeast (*Saccharomyces cerevisiae*), a biological system in which methylation and all the methylation-related machinery are normally absent thus making it a perfect system to study the intrinsic role of DNA methylation on DNA structural and functional properties. With this aim, we expressed the murine DNA Methyl Transferases in *S. cerevisiae* and analyzed the correlation between DNA methylation, nucleosome positioning, gene expression and 3D genome organization. We showed that DNA methylation in our model system followed a conserved pattern, the level of DNA methylation being very low at the 5' end of the gene, and then increasing gradually toward the 3' end. We also observed a correlation between DNA methylation and gene expression: DNA methylation being lower at the TSS and higher at the TTS in highly expressed genes compared to lowly expressed genes and an anti-correlation between DNA methylation and nucleosome positioning. Finally, we showed that DNA methylation tends to increase chromatin condensation, mostly visible in the peri-centromeric region, and decrease DNA flexibility. It also appears to maintain the DNA in its heterochromatin conformation. Taken together, these results provide new insights into the role of DNA methylation and validate our minimal system as a powerful tool to study DNA methylation.

Introduction

DNA methylation is one of the most studied epigenetic marks which introduce major changes in cellular functionality that can have systemic consequences. Thus, the impact of impaired DNA methylation in health is well established (for a review see ([Bird 2002](#))). For example, mutations in DNA Methyltransferase 3b (DNMT3b) are implicated in Immunodeficiency, Centromere instability and Facial anomalies (ICF) syndrome ([Heyn et al. 2012](#)), mutations in DNMT3a are found in acute myeloid leukemia (AML) patients ([Holz-Schietinger et al. 2012](#)) while those in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy ([Winkelmann et al. 2012](#)). Furthermore, DNA methylation plays a key role in development (reviewed in ([Smith and Meissner 2013](#))) and cell differentiation ([Orlanski et al. 2016](#)), and a correct methylation level is crucial for the regulation of parental imprinting and X chromosome inactivation mechanisms ([Miranda and Jones 2007](#)). Finally, changes in DNA methylation patterns have been associated with many different types of cancer in humans ([Kulis et al. 2012](#); [Mayol et al. 2012](#); [Carmona et al. 2014](#); [Subramaniam et al. 2014](#)) and reviewed in ([Heyn and Esteller 2012](#); [Klutstein et al. 2016](#)).

It is commonly stated that DNA methylation in the promoter region of a gene is a hallmark of repression. However, several studies have shown that DNA methylation in the gene body could also affect gene expression, and that the increase in methylation in promoter regions is not always correlated with gene repression, making the effect of DNA methylation on gene expression far more complicated than a simple on/off signal ([Suzuki and Bird 2008](#); [Kulis et al. 2013](#)). Two possible mechanisms might link DNA methylation with gene regulation: i) a direct mechanism involving proteins with methylated DNA binding domains, and ii) an indirect effect related to changes in chromatin structure. For the latter possibility, several hypotheses have been suggested, among them that DNA methylation affects nucleosome positioning, which is known to regulate gene expression by modulating chromatin compaction and DNA accessibility ([Jiang and Pugh 2009](#)). Accurate *in silico* and *in vitro* studies have demonstrated that DNA methylation makes DNA less flexible and less likely to form nucleosomes ([Perez et al. 2012](#); [Portella et al. 2013](#)), but other studies relying on *in vitro* nucleosome reconstitution claim the opposite: *i.e.* that DNA methylation increases the affinity of histones for DNA ([Collings et al. 2013](#)). Also, DNA methylation has been suggested to

promote compaction on a pre-assembled nucleosome ([Choy et al. 2010](#)). *In vivo* studies on mammals and plants are also confusing, with some suggesting that methylation occurs preferentially on nucleosomal DNA ([Cokus et al. 2008](#); [Chodavarapu et al. 2010](#)) and others concluding the opposite ([Felle et al. 2011](#); [Huff and Zilberman 2014](#); [Pedersen et al. 2014](#); [Morselli et al. 2015](#)). This confusion can probably be explained by the high complexity of mammalian genomes and the myriad of factors that can control directly or indirectly nucleosome positioning. For instance, NOME-seq experiments have shown that DNA methylation and nucleosome occupancy were strongly anti-correlated surrounding CTCF sites, but at promoters the correlation seemed to be less clear ([Kelly et al. 2012](#)).

To study the direct effect of DNA methylation on Nucleosome positioning, several groups have used the yeast *Saccharomyces cerevisiae*, taking advantage that this simple eukaryote has neither DNA methylases nor methylated DNA binding domains ([Hu et al. 2009](#)) and ([Bulkowska et al. 2007](#)) have shown that ectopic expression of either DNMT3a and DNMT3L or DNMT3a and DNMT1, could induce DNA methylation in yeast, but the level of methylation achieved in both cases were too low to perform genome wide analysis of the effects of DNA methylation on chromatin structure or gene expression. More recently, Morselli *et al.* achieved a higher level of methylation by expressing DNMT3b at high level and collecting the cells at saturation ([Morselli et al. 2015](#)). Using this approach, they showed that DNA methylation was anti-correlated with nucleosome positioning and that DNMT3b activity correlates positively with H3K36me3 and negatively with H3K4me3. The impact of methylation on the global structure of chromatin is however unknown. In this paper we will present a comprehensive study of the intrinsic impact of DNA methylation in the structure of chromatin, from the nucleosome to the entire chromatin. The study provides us a complete picture of the connection between methylation, chromatin structure and chromatin functional in the absence of specific protein effectors, which allowed us to isolate the direct effect of methylation in modulating DNA physical properties from other effects related to the presence of complex cellular machinery created to recognize specifically methylation signals.

Results

Description of the system

In order to reach a high level of DNA methylation, we expressed the 4 DNMTs simultaneously: yeast cells were transformed with a combination of 4 plasmids each of them expressing one murine DNA Methyl Transferase (DNMT). The *de novo* DNMTs, DNMT3a and 3b, and the maintenance DNMT, DNMT1, were expressed under the control of the *tetO* promoter while DNMT3L, the regulatory DNMT, was expressed under the control of the *Gal1* promoter ([Gari et al. 1997](#)). The expression and stability of the DNMTs were assessed by Western blotting and no sign of protein degradation could be detected, even after 48 hours of induction (Suppl. Figure S1A-D). Expression of the 4 DNMTs slightly affected cell viability and increased the generation time to 7 hours compared to 4 hours for the cells transformed with the 4 empty plasmids grown in the same culture media (Suppl. Figure S2A). Flow cytometry analysis performed on non-synchronized cultures after 24 hours of induction showed a clear increase in the percentage of cells in G2/M (53.7% and 57.2% of cells in two independent transformants) compared to the two control populations transformed with the empty plasmids (40.4% and 41.7%, Suppl. Figure S2B). This result suggests that cells expressing the DNMTs have a slightly longer G2 and/or M phase compared to the control cells. However, differential gene expression analysis (Suppl. Table S1) showed that DNMTs ectopic overexpression did not induce expression of genes normally activated by stress, suggesting that the effects that we observed are not caused by stress but could be a direct effect of DNA methylation

Overall DNA methylation was assessed by HPLC/MS: up to 4.2% of cytosines were methylated after 29 hours of induction in cells collected in early stationary phase. We then determined DNA methylation at single base pair resolution in several independent transformants, using Illumina whole genome bisulfite sequencing (WGBS). This was done both for cells in exponential growth phase synchronized in G1, and for cells at saturation. As illustrated in Table 1, methylated cytosines were almost exclusively found in CpG context, and the pattern of which CpGs were methylated was very reproducible from one sample to another (Figure 1A). To confirm these results, we also performed sequencing using Oxford Nanopore Technology (ONT), which produces much longer read lengths (>10kbp) permitting investigation of repetitive genomic

regions and also allow the correlation of methylation at multiple CpG sites on the same DNA molecule to be assayed. Data were generated for samples with and without the DNMT plasmids for both replicating cells and cells at saturation. Comparisons between the CpG methylation estimates from nanopore and WGBS data show strong correlation with essentially the same pattern of methylation across the genome (Figure S3A-C).

Effect on methylation of the different DNMTs

We tested the functionality of each DNMT in our system by comparing the DNA methylation pattern and levels obtained when only 3 of the 4 DNMTs were expressed (Figure 1A and 1B). This was performed always with replicating cells, synchronized in G1 to make data fully comparable. For all combinations of 3 DNMTs the methylation pattern is broadly the same (Figure 1A) with the same CpGs being methylated. The overall methylation level, however, was 2 to 3 times lower with 3 DNMTs compared to when all 4 DNMTs were expressed, demonstrating that each DNMT is functional in our system (Figure 1B).

While the different DNMT combinations displayed broadly the same pattern of methylated CpGs, when examined in detail some differences emerged. To check whether these differences are due to intrinsic sequence specificity of the different DNMTs, the two bases upstream and downstream of each methylated CpG were considered and logistic regression was used to assess the effect of local sequence context on the methylation rate at CpG sites. This analysis was applied to 6 samples (2 samples with all 4 DNMTs to assess the reproducibility of the results, and the 4 samples each lacking one of the DNMTs). Figure 1C shows that the two replicates with all 4 DNMTs being expressed provide very similar results, and that removing DNMT1 also does not have a strong effect. Removing DNMT3L shows more of an effect, although the comparison with the results from all 4 DNMTs still gives a plot with most points clustered around the diagonal. In contrast, removing either DNMT3a or DNMT3b has a large effect on the sequence context, and this effect is different for the two enzymes with the comparison between the sample without DNMT3a and without DNMT3b showing very little correlation, indicating very different sequence specificities.

Looking in detail at the sequence contexts (The logos in Figure 1C), we cannot see a strong sequence bias with either all 4 DNMTs or when DNMT1 was missing. However, cells lacking DNMT3b show a strong bias for CpGs in the 5' ATCGAG 3' motif (with the percentage of methylated CpGs in an ATCGAG motif being more than six times lower when DNMT3b is removed, compared to the samples with one of the other DNMTs missing or to the sample with all 4 DNMTs induced). Cells lacking DNMT3a also show a sequence bias, in this case towards sequences that are more C rich. Note that the lack of methyl DNA binding protein makes these sequence preferences intrinsic of DNMT and not the result of positioning of the methylases in certain sequences due to auxiliary proteins.

Pattern of methylation

The pattern of methylation obtained in our model is very similar to that observed in higher eukaryotes, with DNA methylation being low at the Transcription Start Site (TSS) increasing toward the end of the genes and reaching a maximum at the Transcription Termination Site (TTS) (Figure 1D), a pattern anticorrelated with H3K4 methylation (Figure 1E). However, even with all 4 DNMTs, the level of methylation rarely exceeds 50% even after 48 hours of induction if the cells are kept constantly dividing (Figure 1D, top panel). We can see that in cells collected at saturation the overall level of methylation is much higher than that of the replicating cells synchronized in G1, while the pattern of methylated CpGs remains essentially identical (Figures 1D, 1F, 1G). To investigate more closely the difference in methylation levels between the replicating cells and those at saturation we fitted a model where for each CpG the methylation level for the replicating cells is a fraction k of the level in the cells at saturation, where k is constant across the genome. The maximum likelihood estimate of k for the two replicate datasets was 0.37 and 0.35, so the methylation levels for the cells at saturation are almost 3 times those for the replicating cells. This difference could be explained by the maintenance machinery not playing properly its role due to the absence of cofactors such as UHRF1 or G9a/GLP or the absence of H3K9 methylation. However, expressing UHRF1 and the H3K9 methylase, SuVar39 in our system did not increase the methylation level in replicating cells (data not shown).

Distinct populations of methylated and non-methylated DNA

The lower levels of methylation in the replicating cells could be due to either (a) each CpG independently have a low probability of being methylated, or (b) because there are distinct methylated and non-methylated populations of DNA. To distinguish between these two scenarios, we examined the long nanopore derived reads, which allow assaying the methylation status at a large number of CpG sites on the same DNA molecule. We can see (Suppl. Figure S3) that the histograms of methylation level for the 2 control samples show an almost identical distribution with a peak at around 0.02, whereas for the methylated sample from cells at saturation the peak is around 0.33. The histogram for the methylated sample from replicating cells has a peak close to the non-methylated controls, but with a long right tail corresponding to a population of reads with a higher proportion of methylated CpGs. To test for evidence that the methylated samples contained distinct populations of reads with different proportions of methylated CpGs, we performed a series of likelihood ratio tests to compare models with distinct read populations to a simple model with homogenous read populations (see Methods). The results are given in Suppl. Figure S4, where we can see strong evidence that the methylated samples contain multiple populations of reads with different CpG methylation rates. Our analysis indicates that around 20% of the reads from the exponential sample are highly methylated (roughly 30% of CpGs on the read being methylated) with the remaining reads having a lower methylation rate of around 8%. For the saturation sample almost 95% of the reads are highly methylated with the remainder having a lower methylation level.

DNA methylation and Nucleosome positioning.

To investigate a possible effect of DNA methylation on nucleosome position, we obtained MNase-Seq data of the samples in stationary phase incubated with all 4 DNMTs or with empty plasmids. The software NucleR ([Flores and Orozco, 2011](#)) and Nucleosome Dynamics ([Buitrago et al. 2019](#)) were used to analyze the MNase-Seq data providing estimates of nucleosome positions as well as an assessment of whether each nucleosome was well positioned or not (fuzzy). As previously reported by ([Morselli et al. 2015](#)), we observed that DNA methylation was anti-correlated with nucleosomes and tended to be accumulated in the linker regions (Figures 2A and 2B). In addition, we observed an increase in the number of fuzzy nucleosomes in the methylated samples

with a 3.87% increase in fuzzy nucleosomes and a 4.01% decrease in well positioned nucleosomes (Suppl. Table S2). In well positioned nucleosomes, DNA methylation was almost absent at the dyad and increased toward the entry and exit points of the nucleosome, while fuzzy nucleosomes had higher methylation levels with a constant level across the nucleosome (compare Figure 2C with Figure 2D). Looking at each strand independently, we observed an asymmetric pattern, the methylation being higher at the 5' extremity, both for the Watson than the Crick strand (Figure 2E). However, we did not observe any periodicity of the methylation signal that was consistent across replicates.

Analysis by Nucleosome Dynamics ([Buitrago et al. 2019](#)) of the crucial region around the promoter (we recently demonstrated that the position of the +1 nucleosome determines most of the nucleosome architecture at the gene body) demonstrates that high level of methylation induces more dynamic and fuzzy nucleosomes (Figure 3A), and narrower Nucleosome Free Regions (NFRs; Figures 3B and C). It is worth noting, that these changes in nucleosome architecture at the crucial promoter region, which are typically considered to be signals of gene inactivation, cannot be explained here by the coordinated effect of methyl-DNA binding proteins coupled to chromatin remodelers.

DNA methylation vs Gene expression

Despite the lack of any mechanism to direct methylation to specific sites, the methylation pattern is quite homogenous throughout the gene only in lowly expressed genes while in the highly expressed ones, DNA methylation is low at the promoter and increase toward the end of the gene, suggesting a link between DNA methylation and gene expression (Figure 4A). A differential expression analysis (Figure 4B) shows that genes which are very lowly methylated do not change their expression level, while high methylation levels lead to important changes in gene activity: while the overall deviation is towards lower expression level, there is a wide distribution of changes across methylated genes. For example, 20 genes are largely down regulated in methylated samples, while 63 are up regulated (Figure 4C). In particular, we see a very strong correlation between gene expression and methylation level for a subset of genes involved in meiosis and that appear to share a common sequence in their regulatory region (Figure 4C, 4D and Suppl. Table S3). Interestingly, this CpG rich motif corresponds to the binding site of Ume6p, a subunit of the histone deacetylase complex

Rpd3p known to repress early meiotic gene expression. It is tempting to hypothesize that methylation of a Ume6p binding site known as URS1, could affect Ume6p binding directly (through changes in direct interactions) or indirectly (through changes in chromatin structure), leading to a deregulation of its target genes. Supporting this hypothesis, we observed that the level of expression of the target genes increases proportionally with the level of methylation of the Ume6p binding site (compare expression and methylation levels in G1 vs Stationary in Suppl. Table S3). Also, for most of these genes, we observe a 5-10bp shift of the -1 or +1 nucleosome (Suppl. Figure S5). In summary, it seems that two physically driven mechanisms: (i) changes in protein-DNA interactions due to the presence of a methyl group at the d(CpG) step and (ii) methylation induced nucleosome rearrangements, coordinate to induce a change in gene activity which would be typically assigned to the effect of methyl DNA binding proteins.

DNA methylation and genome 3D structure

We performed Hi-C experiments in control and methylated populations at saturation to explore the intrinsic effect of DNA methylation in the global chromatin structure. As shown in Figures 5A and 5B (and Suppl. Figures S6A,B), DNA methylation leads globally to a statistically significant decrease in *trans* contacts and increase in *cis* contacts (52% and 53% in the two control samples and 58% and 59% in the two methylated samples (Figure 5C). Looking at the specific regions involved in differential contacts between the control and methylated samples, we identified the regions that significantly gained or lost interactions in the two replicas using the R package diffHic. Regions that gained interactions were preferentially located around the centromeres, increasing the contacts in *cis* between the two arms (Figure 5D), while regions that lost interactions were mostly in *trans*, with 19% of those involving telomeric regions (Figure 5E). Considering only pairs of regions showing a significant gain in interactions, we see that in 25.9% (272 out of 1051) of such pairs at least 1 of the bins contained a rRNA gene. In contrast, when considering all pairs *not* showing a significant change in interactions only 18.9% overlapped with tRNA genes (38677 out of 204435). This enrichment in tRNA is highly significant (Fisher's exact test, $p = 3.7e-8$). Focusing on the regions that showed a reduction in interactions did not yield a significant association with tRNA distribution, though this may be due to the much smaller number of such cases.

To obtain further insights into the effect of DNA methylation on chromatin structure, we modeled the spatial organization of each chromosome using a restraint-based model derived from the interaction counts at 5Kb resolution. Computing the radius of gyration around the centromeres ($\pm 100000\text{bp}$) we confirmed that all chromosomes (except chrI, that is the shortest) are more condensed upon DNA methylation (Figure 6A). We also observed a general tendency decrease of chromosome flexibility with the effect being significant for chromosomes V, IX, XI and XV (Figure 6B and Suppl. Figure S8).

Comparing the significant interactions between the control and methylated samples for each individual chromosome (Suppl. Figure S7), we observed the strongest effects of methylation for chrIII (largest increase in intra-contacts, Figures 6C-F) and for chrXII (largest decrease in intra-chr contacts, Figures 6G-I). The *S. cerevisiae* genome only contains few heterochromatin regions : The two silent loci of the mating type system on chrIII, the rDNA locus on chrXII and the telomeres. Using the nanopore data, we could confirm that both regions were indeed methylated (Suppl. Figures S9A-D). Looking closer at chrIII, we noticed a significant decrease of the distance between the left telomeric region containing the silenced HMLalpha and the more central MATa locus (Figure 6F).

Concerning chrXII, the rDNA repeats localize at the nucleolus and physically segregate the upstream and downstream regions of the chromosome. However, upon methylation, the separation between these two regions in chrXII is much sharper suggesting a structural modification of the rDNA locus.

Discussion

The first striking result of this study is that we could induce, *in vivo*, a specific pattern of DNA methylation reproducible and similar to that of mammals in an organism *a priori* deprived of any DNA methylation machinery. This implies that the DNMTs are sufficient to define many aspects of mammalian chromatin structure in a system natively lacking any machinery capable of recognizing methylation patterns. Methylation leads to some phenotypic changes in the cells, the most visible one being a longer than normal G2/M phase. However, such changes are moderate and an organism not prepared to have methylated DNA appears to tolerate well a non-negligible amount of methylation in its genome.

The level of DNA methylation that could be obtained in exponentially growing cells did not go over a certain threshold even if the time of induction is increased. However, this level is more than 3 times higher when cells are in stationary phase, suggesting that the overall low level of DNA methylation in dividing cells is due to the poor ability of DNMT1 to act as a maintenance transferase. This is likely to be related to the absence of cofactors like Ring-finger domain UHRF1 or even the lack of H3K9 methylation in yeast, which might reduce specificity of DNMT1 for hemi-methylated DNA.

Our synthetic model system helped us to highlight some previously unknown intrinsic sequence specificity for the methyl-transferases. For example, little sequence specificity is found for DNMT1 and DNMT3L, while AAA differences in sequence specificity are found for DNMT3a and DNMT3b. Thus, cells lacking DNMT3b show a strong bias for methylation of CpG in 5' ATCGAG 3' motif, while cells lacking DNMT3a seems to have a bias toward CpG sequences embedded in C rich environments.

Methylation is preferentially located between nucleosomes, and when it occurs in nucleosome-occupied regions (in the basal non-methylated control), it can be associated with significant alterations in nucleosome positioning reflected by an increase in nucleosome fuzziness. The fact that methylated DNA is less frequent at nucleosomes, confirms our previous *in silico* and *in vitro* models ([Perez et al. 2012](#); [Portella et al. 2013](#)), but does not rule out the possibility that methylated-DNA binding domains might stabilize the presence of methylated CpG in nucleosomes, leading to a situation of “loading-spring” which might facilitate fast and nucleosome reorganization upon release of the stabilizing protein. There is, however, no question that methylation and nucleosome position are intrinsically anti-correlated.

However, when comparing nucleosome occupancy for highly and lowly methylated promoters, we observed that the width of the NFR region tends to be narrower when the promoter is more methylated. This could be explained by the fact that wider NFR are occupied by transcription factors or remodelers that could prevent the methylation machinery to access the DNA.

In mammalian cells, the relationship between methylation and gene expression is complex, with high levels of gene expression often associated with low promoter methylation but elevated gene body methylation, and the causality relationships is not always clear. We used our system to check whether or not we could find a general correlation between DNA methylation and gene expression. We observed that despite

the lack of specific methylated DNA binding domains in yeast, highly and lowly expressed genes have quite different methylation profiles, with much higher levels of methylation near (± 850 bp) the TSS of silent genes while highly active genes have much higher methylation levels at the TTS. In the absence of specific proteins modulating this profound difference, we can speculate that nucleosome positioning is one of the main factors responsible for this differential behavior, which suggests that methylation and nucleosome positioning might act in concert in the regulation of gene function in mammals, adding an extra layer of control of gene expression.

However, our results showed that methylation can also directly affect the binding of a transcription factor and we think this is what is happening for the early meiotic genes whose expression dramatically increases in response to methylation. In that case, we suspect the methylation to affect the binding of the histone deacetylase complex Rpd3p that acts as a repressor.

There is yet another level of regulation which relies on the chromatin structure and we were able to use our system to show that DNA methylation had a general impact on 3D genome organization. Indeed, even if the genome retains its characteristic Rabl configuration previously observed in exponential as well as in quiescent cells ([Duan et al. 2010](#); [Rutledge et al. 2015](#)) upon methylation, we observed a significant increase of intra-chromosomal contacts and a significant decrease of inter-chromosomal contacts, of which a significant proportion involves regions containing one or several tRNA genes. This could be explained by the putative role of TFIIC on the 3D genome organization suggested by ([Noma et al. 2006](#)).

One major observation is the condensation of the centromeric region as illustrated by the gain of interactions between pericentromeric regions and by the smaller radius of gyration of this region in methylated cells compared to our control cells. Also, we observed a tendency for the chromosome to be less flexible, which is in agreement with the effect of DNA methylation described *in vitro* ([Perez et al. 2012](#)).

As DNA methylation is associated with silencing and with heterochromatin, we focused on the three heterochromatin regions of the *S. cerevisiae* genome: the telomeres, the mating type locus and the rDNA locus. First, we observed a general loss of interactions of the telomeric regions. This could be explained by a lower flexibility of the chromosomes as previously mentioned. In the case of the mating type locus, we

observed that the silenced locus HML α was closer to the Mat locus, a conformation expected for exponentially growing MATa cells ([Belton et al. 2015](#)) but less frequent in stationary phase (reflected by the increase of interaction between the two telomeres 3L and 3R in ([Rutledge et al. 2015](#))). Finally, the separation between the two regions of chrXII divided by the rDNA locus, is much stronger upon DNA methylation suggesting that methylation of the rDNA increases its rigidity and prevent any interactions between the two contiguous regions.

Very interestingly, when ([Rutledge et al. 2015](#)) compared the 3D genome structure in exponential and quiescent cells, he reported an increase of interactions between telomeres and an increase in the long range intra-chromosomal interactions in chrXII in quiescent cells (it should be noted that in the later case, intra-chromosomal interaction between the two regions separated by the rDNA locus are only mildly increased compared to the one within each region). Our results are showing that those interactions are prevented upon DNA methylation, suggesting that DNA methylation could freeze the heterochromatin structure in the conformation it had originally (before the methylation was induced), i.e. in the exponential conformation.

Materials and Methods

Plasmid construction

pYADE4 yeast plasmids encoding full length DNMT1 and DNMT3a with modified sequences around the translation start sites were kindly provided by Dr Jan Fronck, pYES3/CT encoding DNMT3b was provided by Dr Shen Li ([Bulkowska et al. 2007](#); [Shen et al. 2010](#)). DNMT3L cloned into pYES3/CT to produce a Nterminal FLAG tagged DNMT3L was provided by Dr Jia-Lei Hu ([Hu et al. 2009](#)).

pCM188 (marker cgURA3) and pCM185 (marker cgHIS or cgLEU), centromeric vectors which differ for the number of Tet operators (respectively 2 and 7;([Gari et al. 1997](#))) were kindly provided by Dr Jessie Colin.

*Sma*I restriction fragment (from pYADE4-DNMT1) containing full length DNMT1 cDNA was inserted at *Pme*I site of pCM185 (LEU) to give pCM185(LEU)-DNMT1.

*Bam*HI-*Mlu*I restriction fragment (from pYADE4-DNMT3a) containing full length cDNA from DNMT3a was ligated to pCM185 (HIS) linearized with *Bam*HI and *Mlu*I to give pCM185(HIS)-DNMT3a.

*Bam*HI-*Not*I restriction fragment from pYES3/CT-DNMT3b containing full length DNMT3b was ligated to pCM188 (URA) linearized with *Bam*HI and *Not*I to obtain pCM188 (URA)-DNMT3b.

Yeast strains and culture conditions

Strain YPH499 (*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1*) was transformed with 2, 3 or 4 expression plasmids by the standard lithium acetate procedure. Transformants were selected on plates of appropriate selective medium with 2% Raffinose and 10μg/ml doxycycline to repress any expression.

Selected transformants (2 to 4 transformants per combination of plasmids) were grown on selective liquid medium with 2% Raffinose and 10μg/ml doxycycline up to OD₆₀₀=0.5. Then, yeast cells were spun 10min at 1000x g, washed twice with sterilized water, and resuspended into selective media with 1% Raffinose and 2% Galactose without doxycycline to allow expression of DNMTs. For experiments on synchronized cells, cells were treated with alpha-factor (3μM final) for 4 hours to synchronise cells in G1 or with Nocodazole to synchronize cells in G2.

After different times of induction, cells were collected and treated for subsequent experiments: protein extraction for western blotting, gDNA extraction for whole genome bisulfite sequencing, RNA extraction for RNA-sequencing or Semi-intact cell preparation for Mnase digestion and nucleosome mapping)

Flow cytometry analysis

0.5 ml of culture (OD₆₀₀=0.6-0.8) were collected and centrifuged for 5 mn at 1000 g at RT. Pellet were washed twice with 1x ice-cold PBS and resuspended in 50 μl of 1x ice-cold PBS. 20 μl of cells were fixed with 1 ml of 70 % EtOH overnight at 4 °C. Samples were washed with 1x Saline Sodium Citrate buffer (SSC; 150 mM NaCl, 15 mM Nacitrate, pH 7.8 for 20x SSC). The pellet was resuspended in 0.5 ml of 1x SSC, treated with 0.5 mg/ml RNase A (Roche) for 1.5 h and then with 0.5 mg/ml Proteinase K (Roche) for another 1.5 h at 50 °C. After incubation, cells were briefly sonicated for 10 mn, medium

potency, by using the *Bioruptor* system (intervals of 10 s on-20 s off). 250 µl of the cells were added in 0,5 ml of 1x SSC containing 1 µM Sytox Green (Sigma) and were incubated 10-20mn in the dark (room temperature) before analyzing the DNA content using a Beckam Coulter Gallios™ flow cytometer.

HPLC/MS

HPLC/MS/MS analysis was based on the protocol described ([Friso et al. 2002](#)). A Kinetex 2.6 µm HILIC 100A column (150 mm × 4.6 mm) (Phenomenex) and a Acquity UPLC system (Waters Corp., Milford, MA, USA) coupled to a mass spectrometer API 3000™ (AB Sciex, Foster City, CA, USA) triple quadrupole working in MRM(multiple reaction monitoring) method in positive mode. Two eluents were used: eluent A2 (Acetonitrile) and eluent B1 (0.1 M ammonium formiate adjusted at pH 3.2) with a isocratic gradient 8 min of total running time at 90 % A and 10 % B for the nucleosides elution. The separation was performed in a flow of 1400 µl min⁻¹, with 10 µl injection volume and two replicates each, totaling two biological replicates and two technical replicates of each sample. The standard nucleosides cytosine and methyl-cytosine (Sigma) were diluted in HCl 0.01N and stored at -20 °C. The m/z transitions from 112 to 95 (cytosine) and from 126 to 81 (methyl cytosine) were chosen for MRM experiments. The peak area obtained was analyzed by Analyst 1.4.2 (AB Sciex). Quantification (%) was performed according to 5mdC concentration divided by 5mdC concentration plus dC concentration multiplied by 100.

Western Blot

Proteins were extracted by resuspending the pellet of cells from a 20ml cultures at OD600=1 in 400µl of RIPA buffer (50mM Tris pH7.5, 150mM NaCl, 1% NP40, 0.5% NaDeoxycholate, 0.1% SDS) containing 1mM PMSF and protease inhibitors (cOmplete ULTRA Tablets, Mini, EASYpack, Roche). 400µl of glass beads were added and samples were processed using FastPrep (MP) for 3 times for 20sec pulses @4.5m/s. After centrifugation 5min at 5000rpm, supernatant were recovered and quantified by bradford. 20 µg of protein were loaded on 6 or 8 % acrylamide gel and subjected to PAGE, proteins were then transferred onto an immobilon membrane (millipore) for subsequent hybridization with anti-DNMT1 (ref ab87654, Abcam), anti-DNMT3a (ab2850, Abcam), anti-DNMT3b (ab122932, Abcam) or anti-Flag (F7425, Sigma) antibody overnight followed by secondary antibody anti Rabbit (Goat)-HRP conjugated

(65-6120, Invitrogen). The signal was revealed using ECL™ prime WB detection reagent (Amersham, GE Healthcare).

Semi-intact Yeast cell preparation

Semi-intact cells were prepared as previously described ([Schlenstedt et al. 1993](#)). Briefly, cells were grown at 30°C in 300 ml YPD to $\approx 1 \times 10^7$ cells/ml. For each 250 ml of cells (10^7 cells/ml), semi-intact cells were prepared as follows. Cells were collected by centrifugation (700 g, 7 min, RT), resuspended in 25 ml 100 mM Pipes, pH 9.4, 10 mM DTT, incubated with gentle agitation at 30°C for 10 min, and collected by centrifugation (1,000 g, 5 min, RT). Cells were resuspended in 6 ml YP, 0.2% glucose, 50 mM KPO₄, pH 7.5, 0.6 M sorbitol. 10u zymolase was added, and the suspension was incubated with gentle shaking 30°C for 30 min. Spheroplasting was monitored by light microscopy. Great care was taken not to overdigest cells to avoid lysis. Spheroplasts were collected by centrifugation at 1,000 g for 5 min at RT, re-suspended with a plastic pipette in 40 ml YE 1% glucose, 0.7 M sorbitol, and incubated with gentle shaking at 30°C for 20 min. Spheroplasts were collected by centrifugation (1,000 g, 5 min, RT) and washed twice at 4°C with cold permeabilization buffer (20 mM Pipes-KOH, pH 6.8, 150 mM K-Acetate, 2 mM Mg-Acetate, 0.4 M sorbitol). The final pellet was resuspended in 1ml cold permeabilization buffer containing 10%(v/v)DMSO. 100µl aliquots were placed in 1.5 ml microfuge tubes and frozen slowly above liquid N₂ and stored at -80°C.

ChIP-seq

Yeast strains were grown into the appropriate selective media with 2% galactose and 1% raffinose to stationary phase (OD₆₀₀ similar to 8) and diluted to OD₆₀₀=1 in 50 ml of the media, next cells were crosslinked 20 min with 1% of formaldehyde followed by a 15 min incubation with 125 mM of glycine. After crosslinking, spheroplast were isolated as described before using the zymolase enzyme and resuspended in 0.3 ml of lysis buffer (50 mM Hepes-KOH at pH 7.2, 140 mM NaCl, 1 mM EDTA, 0.1% Deoxycholic acid sodium salt and 1% Triton X-100) containing a cocktail of protease inhibitors (Roche, 04693159001). An equal volume of glass beads (0.5-mm diameter) was added, and the spheroplast were broken using a bead-beater (FastPrep-24, Biomedicals). Glass beads were then removed and the lysate was transferred to a Sorenson tubes to digest the

chromatin into fragments of 300 nucleotides using the Bioruptor Pico (30 cycles, 30"on/30"off). The whole extract was clarified by centrifugation for 10 min at $5000 \times g$ at 4°C and an aliquot was taken as input. In parallel, 50 µl of Dynabeads M-280 Sheep anti Rabbit IgG (Thermo Fisher) per sample were washed twice with PBS+5mg/ml of BSA and incubated at least for 12 hours at 4°C with 2.5 µg of the primary antibody (HA, H3K4me, H3K4me3 from Abcam). Next, beads were washed again with PBS+5mg/ml of BSA and resuspended in 30 µl/sample of PBS-BSA 5mg/ml. Extracts were then incubated 2h at 4°C with the Dynabeads, previously conjugated with the primary antibody, and then washed two times with lysis buffer, two times with the lysis buffer supplemented with 360mM NaCl, 2 times with the wash buffer (0.5% Deoxycholic acid sodium salt, 10mM TRIS pH8, 250 mM LiCl, 0.5% NP-40, 1 mM EDTA) and one time with TE (10Mm Tris-HCl pH 7.5, 1mM EDTA). Then Dynabeads were eluted with 80 µl of Elution Buffer (50 mM TRIS, 10 mM of EDTA, 1% SDS) followed by an incubation of 10' at 65°C with continuous mixing. Crosslinking was reverted by leaving the samples at least for 12 hours at 65°C. Finally proteins were digested with 0.80 mg/ml per sample of proteinase K for 2h at 37°C and DNA was purified by phenol-chloroform and chloroform extractions and ethanol precipitation.

MNase-seq

0.4×10^9 semi-intact cells were digested with micrococcal Nuclease (MNase), 1.5 unit at 37°C for 30min with 3mM CaCl₂. The reactions were stopped by addition of EDTA to a final concentration of 0.02 M and subsequently incubated with RNase A (0.1 mg) for 4h at 37°C and further treated with Proteinase K at 37°C o/n. DNA was purified using phenol-chloroform extraction and concentrated by ethanol precipitation.

The percentage of mononucleosomal DNA fragments was examined by 2% agarose gels. Furthermore, the integrity and size distribution of digested fragments were determined using the microfluidics-based platform Bioanalyzer (Agilent) prior to sample preparations and sequencing following Illumina standard protocol.

Nucleosome calling

MNase-seq paired-end reads were mapped to yeast genome (SacCer3, Apr. 2011) using Bowtie ([Langmead et al. 2009](#)) aligner, allowing a maximum of 2 mismatches and maximum insert size of 500 bp. Output BAM files were imported in R ([Team 2011](#)) and

quality control was performed with htSeqTools package to remove PCR artifacts ([Planet et al. 2012](#)). Filtered reads were processed with nucleR package ([Flores and Orozco 2011](#)) as follows: mapped fragments were trimmed to 50bp maintaining the original center and transformed to reads per million. Then, noise was filtered through Fast Fourier Transform, keeping 2% of the principal components, and peak calling was performed using the parameters: peak width 147 bp, peak detection threshold 35%, maximum overlap of 80 bp, dyad length 50 bp. Nucleosome calls were considered well-positioned when nucleR peak width score and height score were higher than 0.6 and 0.4, respectively, and fuzzy otherwise.

Nucleosome Dynamics

NucDyn R package ([Buitrago et al. 2019](#)) was used to find changes in nucleosome organization between control and methylation induced samples. P-values quantifying the nucleosome change were obtained running NucDyn with the following parameters: maximum difference of 70, maximum length of 140, minimum number of reads to report a shift of 3, shifts threshold of 0.1, indels minimum number of reads to report evictions and inclusions (indels) of 3, indels threshold of 0.05.

Whole-genome bisulfite sequencing (WGBS)

WGBS was performed following the procedure outlined in ([Kulis et al. 2012](#)). Briefly, genomic DNA (1-2µg) was spiked with unmethylated λ DNA (5 ng of λ DNA per µg of genomic DNA) (Promega). The DNA was sheared by sonication to 50-500 bp using a Covaris E220 and fragments of size 150-300 bp were selected using AMPure XP beads (Agencourt Bioscience Corp.). Genomic DNA libraries were constructed using the Illumina TruSeq Sample Preparation kit (Illumina Inc.) following the Illumina standard protocol: end repair was performed on the DNA fragments, an adenine was added to the 3' extremities of the fragments and Illumina TruSeq adapters were ligated at each extremity. After adaptor ligation, the DNA was treated with sodium bisulfite using the EpiTaxy Bisulfite kit (Qiagen) following the manufacturer's instructions for formalin-fixed and paraffin-embedded (FFPE) tissue samples. Two rounds of bisulfite conversion were performed to assure a conversion rate of over 99%. Enrichment for adaptor-ligated DNA was carried out through 7 PCR cycles using the PfuTurboCx Hotstart DNA polymerase (Stratagene). Library quality was monitored using the Agilent 2100 BioAnalyzer (Agilent), and the concentration of viable sequencing fragments (molecules

carrying adaptors at both extremities) estimated using quantitative PCR with the library quantification kit from KAPA Biosystem. Paired-end DNA sequencing (2x100bp) was then performed using the Illumina Hi-Seq 2000.

Read mapping and estimation of cytosine methylation levels

The WGBS reads were processed using the gemBS pipeline v3.0 (REF) using as reference *S. cerevisiae* S288c. Reads with MAPQ scores < 20 and read pairs mapping to the same start and end points on the genome were filtered out after the alignment step. The first 5 bases from each read were trimmed before the variant and methylation calling step to avoid artifacts due to end repair. For each sample, CpG sites were selected where both bases were called with a Phred score of at least 20, corresponding to an estimated genotype error level of $\leq 1\%$. Sites with >500x coverage depth were excluded to avoid centromeric/telomeric repetitive regions. CpGs were considered methylated when the number of mapped reads was larger than 10 and the estimated methylation percentage was above 0.1.

Nanopore sequencing

Suspensions of spheroplasts from methylated and control *S. cerevisiae* strains were loaded on Sage Science gel cassettes to perform lysis under electrophoretic conditions. DNA content in each sample was estimated by the cell count. A number of spheroplasts equivalent to 10µg of genomic DNA were resuspended in 70 µl of HLS Suspension buffer (Sage Science, Mammalian white Blood cell suspension kit, #CEL-MWB1) and loaded on the gel cassettes (Sage Science, SageHLS HMW DNA extraction kit #HEX-0012).

The custom Sage HLS (Sage Science) protocol used (Extraction Collection DC55V 1h15m) was accommodated for the yeast small chromosome sizes. This custom protocol did not include a DNA fragmentation step. In brief, during the extraction step, the High Molecular Weight (HMW) yeast gDNA was bound in agarose while the solubilised and degraded proteins and other contaminants were kept in solution. The Sage Science Buffer A was used as a lysis buffer for this step. In the last step of the protocol, the HMW DNA was retrieved from the gel through an automated elution process that was optimized to elute all the yeast chromosomes in the elution module number 2 of the cassette.

Elution modules 1, 2, 3 & 4 were selected for the library preparation of the control and methylated *S. cerevisiae* samples. For each condition, the selected elution modules were pooled, purified with 1-fold excess of Agencourt AMPure XP beads (Beckman Coulter, A63882) and eluted in water. Two barcoded libraries containing both type of samples were prepared using the Oxford Nanopore Ligation sequencing kit (ONT, SQK-LSK109) combined with the Oxford Nanopore Native Barcoding Expansion kit (EXP-NBD103 1D) following manufacturer's instructions.

After connecting the flows cells to the MinION Mk1b device, the MinKNOW interface QC (Oxford Nanopore Technologies) was run in order to assess the flow cell quality. Once the priming of the flow cell was finished, from 200ng to 600ng of the final barcoded library was loaded into R9.4.1 FLO-MIN106 or FLO-MIN106D flow cells and the sequencing data were collected during 48 hours. The quality parameters of the sequencing runs were further monitored by the MinKNOW platform in real time. The MinKNOW versions used was 1.15.4. The basecalling was performed using Guppy 2.3.7. Reads were mapped using minimap2 2.9-r720, and CpG methylation was called using nanopolish 0.11.0.

mRNA library preparation and sequencing

The RNASeq libraries were prepared from total RNA (extracted by the standard hot phenol protocol) using the TruSeq™ RNA Sample Prep Kit v2 (Illumina Inc.,) according to manufacturer's specifications. Briefly, after poly-A based mRNA enrichment with oligo-dT magnetic beads and 0.5µg of total RNA as the input material the mRNA was fragmented (resulting RNA fragment size was 80-250nt, with the major peak at 130nt). After first and second strand cDNA synthesis the double stranded cDNA was end-repaired, 3'adenylated and the 3'-"T" nucleotide of the adapter was used for the Illumina barcoded adapters ligation. The ligation product was enriched by 15 cycles of PCR.

Each library was sequenced using TruSeq SBS Kit v3-HS, in paired-end mode with a read length of 2x76bp. We generated over 20 million paired-end reads for each sample in a fraction of a sequencing lane on HiSeq2000 (Illumina, Inc) following the manufacturer's protocol. Images analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ sequence files by CASAVA.

Mapping and quantification

Around 95% of the reads were mapped against the reference genome (*S.cerevisiae* release 74 + artificial plasmids) with the GEM software (v1.7.0) ([Marco-Sola et al. 2012](#)) allowing for split maps. As expected, most of the reads mapped to exonic regions (92%). Genes were quantified using Flux-Capacitor (v1.6.1) ([Montgomery et al. 2010](#)) and normalized by the TMM method of the edgeR software ([Robinson and Oshlack 2010](#)).

Genes were ranked by their normalized expression values in samples V126-V131 and we selected those with the 10% lower and 10% upper values as highly and lowly expressed, respectively.

Genomic annotation

Data was annotated from the UCSC gene track that contains 6692 genes. We discarded genes that are described as “Putative” or “Dubious” and genes located in the mitochondrial chromosome. We used gene lengths to normalize methylation proportions, nucleosome coverages and CpG density partitioning each gene in 137 bins (each bin has on average 10 bp since the mean length of yeast genes is 1369 bp).

DNMT specificity analysis

We extracted two bases downstream and upstream from each CpG (having at least ten WGBS reads mapped) and trained a logistic regression model (using R) for the number of converted and non-converted Cs, using the extracted motifs as predictors for each WGBS sample (samples removing one of the DNMTs, T859, T860, T861 and T869; and two samples with the four DNMTs, T862 and T863). We computed for each sample the effect of each motif and its standard deviation, and used it to determine those with a significant effect on methylation level (estimated effect above two standard deviations). We found motifs specific for each sample lacking one of the DNMTs (motifs with significant effect in the sample removing one DNMT but not significant in the sample with all DNMTs) and compared their relative frequencies in all samples.

Hi-C data processing and normalization

We processed Hi-C data using TADbit ([Serra et al. 2017](#)) (<https://github.com/3DGenomes/tadbit>) for quality control, mapping and filtering.

First, quality control was performed with the FastQC protocol implementation in TADbit. Then, reads were mapped to the reference yeast genome (sacCer3, Apr. 2011) with a fragment-based strategy. Afterwards, non-informative contacts (self-circle, dangling-end, error, duplicated and random breaks) identified by TADbit were filtered-out, obtaining 32–37 million valid interactions per experiment. Off-target contacts (neither end of the read mapped to one of the capture regions) were also discarded (full details of the number of excluded reads are given in Suppl. Table S4). Finally, contact matrices were created from valid reads at 5 kb resolution with the corresponding TADbit module, and low frequency bins were removed.

Contact matrices were transformed to *.hic* format for visualization in Juicebox ([Durand et al. 2016](#)) using the *pre* command, and normalized with the Balanced method ([Rao et al. 2014](#)).

Differential Hi-C analysis was performed using the R/Bioconductor package diffHic ([Lun, 2015](#)). The mapped Hi-C data were filtered and the differential interaction analysis between the control and methylated samples (using the two replicates for each treatment) was performed using the procedure recommended in the diffHic manual.

Hi-C-based chromatin 3D structure

High resolution Hi-C data at 5 kb was used to obtain the 3D structure, conformation and dynamics of entire yeast chromosomes. The Hi-C technique provides interaction contacts between DNA fragments. The interaction counts or frequencies between two *loci* i and j (f_{ij}) can be converted to spatial 3D distances between those *loci* (d_{ij}) by an inverse relationship (equation 1),

$$d_{ij} = \gamma / f_{ij}^{\alpha} \quad (1)$$

where γ represents the scale of the structure and is usually taken to match experimental distances between selected genomic regions, and the precise value of α depends on the organism under study, the genomic distance, and the resolution of the Hi-C map and needs to be fitted ([Adhikari et al. 2016](#); [Zhang et al. 2013](#); [Varoquaux et al. 2014](#)).

Since Hi-C interaction counts are known to present several biases, such as mappability of fragments, GC content, and fragment length, they were normalized using iterative correction and eigenvector decomposition ([Imakaev et al. 2012](#)). Finally, the output of

the conversion procedure was a matrix containing equilibrium distances (r_0) for the different interacting *loci*. To remove the background noise, a cutoff of two times the median of all trans contacts (i.e., between *different* chromosomes) was applied to the Hi-C contact map to define *interacting regions*.

The chromosome model was built as a chain of beads, each bead representing a genomic region that corresponds to a bin from the Hi-C map. Spatial equilibrium distances were obtained from equation 1 as explained above. The distances between interacting beads (r) were restrained near their equilibrium length during the simulations by penalizing with a harmonic potential (equation 2) when approaching at shorter distances or moving away at longer distances than the equilibrium. A tolerance of one bead radius was applied, thus resulting in a flat-welled parabola potential.

$$E = k(r - r_0)^2 \quad (2)$$

To ensure proper connectivity of the fiber, consecutive beads were also bound by a harmonic potential but with a force constant five orders of magnitude stronger than that applied to interacting non-consecutive beads. An excluded volume was defined for each bead by a standard Lennard-Jones potential with equilibrium distance equal to one bead radius and a soft energy well. Additional repulsive restraints were added for *non interacting* beads, forced to remain at a distance longer than the maximum equilibrium distance obtained from equation 1. The initial structure of the chromosome fiber was varied between an extended conformation and a random localization of initially unbound beads in different replicas. The system was allowed to sample the conformational space using pmemd simulation engine for GPU from Amber 18 package. Different conformations of the fibers were determined by attraction and repulsion forces arising from the distance restraints between beads.

In the end, an ensemble of structures was obtained by minimizing the number of experimental restraint violations (equilibrium distances input). A method yielding a population of structures with different conformations was chosen since Hi-C maps are derived from population of cells with variable chromatin structure.

Data Access

Raw and processed WGBS, RNA-seq and MNase-seq and Hi-C data have been submitted to the European Nucleotide Archive (ENA) under accession number XXXX (not yet available).

Acknowledgments

We thank all the members of the EBL, and especially Antonio Rodriguez Campos for helpful discussions and Nuria Villegas Forn for technical support. We acknowledge the IRB Biostatistic Core Facility for their help with the ChIP-seq analysis. We also want to thank Ron Schuyler, Mike Goodstadt, François Serra and David Castillo from the CRG-CNAG, for fruitful discussions. We are grateful to Dr. Jessie Colin for providing various expression vectors and to Dr Jan Fronck, Dr Shen Li and Dr Jia-Lei Hu for providing DNMT1, DNMT3a, DNMT3b and DNMT3L cDNA.

This work has been supported by the Spanish Ministry of Science (BIO2012-32868), the Catalan SGR, the Instituto Nacional de Bioinformática, and the European Research Council (ERC_SimDNA).

References

- Adhikari B, Trieu T, Cheng J. 2016. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC Genomics* **17**: 886.
- Bashtrykov P, Jankevicius G, Jurkowska RZ, Ragozin S, Jeltsch A. 2014. The UHRF1 Protein Stimulates the Activity and Specificity of the Maintenance DNA Methyltransferase DNMT1 by an Allosteric Mechanism. *J Biol Chem* **289**: 4106-4115.
- Belton JM, Lajoie BR, Audibert S, Cantaloube S, Lassadi I, Goiffon I, Bau D, Marti-Renom MA, Bystricky K, Dekker J. 2015. The Conformation of Yeast Chromosome III Is Mating Type Dependent and Controlled by the Recombination Enhancer. *Cell reports* **13**: 1855-1867.
- Berkyurek AC, Suetake I, Arita K, Takeshita K, Nakagawa A, Shirakawa M, Tajima S. 2014. The DNA methyltransferase Dnmt1 directly interacts with the SET and RING finger-associated (SRA) domain of the multifunctional protein Uhrf1 to facilitate accession of the catalytic center to hemi-methylated DNA. *J Biol Chem* **289**: 379-386.
- Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6-21.
- Buitrago D, Codó L, Illa R, de Jorge P, Battistini F, Flores O, Bayarri G, Del Pino M, Heath S, Hospital A et al. 2019. Nucleosome Dynamics: A new tool for the dynamic analysis of nucleosome positioning. *Nucleic Acids Res.*
- Bulkowska U, Ishikawa T, Kurlandzka A, Trzcinska-Danielewicz J, Derlacz R, Fronk J. 2007. Expression of murine DNA methyltransferases Dnmt1 and Dnmt3a in the yeast *Saccharomyces cerevisiae*. *Yeast* **24**: 871-882.

- Carmona FJ, Davalos V, Vidal E, Gomez A, Heyn H, Hashimoto Y, Vizoso M, Martinez-Cardus A, Sayols S, Ferreira HJ et al. 2014. A comprehensive DNA methylation profile of epithelial-to-mesenchymal transition. *Cancer research* **74**: 5608-5619.
- Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ et al. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**: 388-392.
- Choy JS, Wei S, Lee JY, Tan S, Chu S, Lee TH. 2010. DNA methylation increases nucleosome compaction and rigidity. *J Am Chem Soc* **132**: 1782-1783.
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**: 215-219.
- Collings CK, Waddell PJ, Anderson JN. 2013. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res* **41**: 2918-2931.
- Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* **465**: 363-367.
- Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**: 99-101.
- Felle M, Hoffmeister H, Rothhammer J, Fuchs A, Exler JH, Langst G. 2011. Nucleosomes protect DNA from DNA methylation in vivo and in vitro. *Nucleic Acids Res* **39**: 6956-6969.
- Ferry L, Fournier A, Tsusaka T, Adelmant G, Shimazu T, Matano S, Kirsh O, Amouroux R, Dohmae N, Suzuki T et al. 2017. Methylation of DNA Ligase 1 by G9a/GLP Recruits UHRF1 to Replicating DNA and Regulates DNA Methylation. *Molecular cell* **67**: 550-565 e555.
- Flores O, Orozco M. 2011. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**: 2149-2150.
- Friso S, Choi SW, Dolnikowski GG, Selhub J. 2002. A method to assess genomic DNA methylation using high-performance liquid chromatography/electrospray ionization mass spectrometry. *Analytical chemistry* **74**: 4526-4531.
- Gari E, Piedrafito L, Aldea M, Herrero E. 1997. A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in *Saccharomyces cerevisiae*. *Yeast* **13**: 837-848.
- Heyn H, Esteller M. 2012. DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* **13**: 679-692.
- Heyn H, Vidal E, Sayols S, Sanchez-Mut JV, Moran S, Medina I, Sandoval J, Simo-Riudalbas L, Szczesna K, Huertas D et al. 2012. Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient. *Epigenetics : official journal of the DNA Methylation Society* **7**: 542-550.
- Holz-Schietinger C, Matje DM, Reich NO. 2012. Mutations in DNA methyltransferase (DNMT3A) observed in acute myeloid leukemia patients disrupt processive methylation. *J Biol Chem* **287**: 30941-30951.
- Hu JL, Zhou BO, Zhang RR, Zhang KL, Zhou JQ, Xu GL. 2009. The N-terminus of histone H3 is required for de novo DNA methylation in chromatin. *Proc Natl Acad Sci U S A* **106**: 22187-22192.
- Huff JT, Zilberman D. 2014. Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. *Cell* **156**: 1286-1297.
- Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999-1003.

- Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**: 161-172.
- Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**: 2497-2506.
- Klutstein M, Nejman D, Greenfield R, Cedar H. 2016. DNA Methylation in Cancer and Aging. *Cancer research* **76**: 3446-3450.
- Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, Martinez-Trillos A, Castellano G, Brun-Heath I, Pinyol M et al. 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* **44**: 1236-1242.
- Kulis M, Queiros AC, Beekman R, Martin-Subero JI. 2013. Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochimica et biophysica acta* **1829**: 1161-1174.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Liu X, Gao Q, Li P, Zhao Q, Zhang J, Li J, Koseki H, Wong J. 2013. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nature communications* **4**: 1563.
- Lun AT, Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**: 258.
- Marco-Sola S, Sammeth M, Guigo R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**: 1185-1188.
- Mayol G, Martin-Subero JI, Rios J, Queiros A, Kulis M, Sunol M, Esteller M, Gomez S, Garcia I, de Torres C et al. 2012. DNA hypomethylation affects cancer-related biological functions and genes relevant in neuroblastoma pathogenesis. *PLoS One* **7**: e48401.
- Miranda TB, Jones PA. 2007. DNA methylation: the nuts and bolts of repression. *Journal of cellular physiology* **213**: 384-390.
- Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773-777.
- Morselli M, Pastor WA, Montanini B, Nee K, Ferrari R, Fu K, Bonora G, Rubbi L, Clark AT, Ottonello S et al. 2015. In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. *Elife* **4**: e06205.
- Noma K, Cam HP, Maraia RJ, Grewal SI. 2006. A role for TFIIC transcription factor complex in genome organization. *Cell* **125**: 859-872.
- Orlanski S, Labi V, Reizel Y, Spiro A, Lichtenstein M, Levin-Klein R, Koralov SB, Skversky Y, Rajewsky K, Cedar H et al. 2016. Tissue-specific DNA demethylation is required for proper B-cell differentiation and function. *Proc Natl Acad Sci U S A* **113**: 5018-5023.
- Pedersen JS, Valen E, Velazquez AM, Parker BJ, Rasmussen M, Lindgreen S, Lilje B, Tobin DJ, Kelly TK, Vang S et al. 2014. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res* **24**: 454-466.
- Perez A, Castellazzi CL, Battistini F, Collinet K, Flores O, Deniz O, Ruiz ML, Torrents D, Eritja R, Soler-Lopez M et al. 2012. Impact of methylation on the physical properties of DNA. *Biophysical journal* **102**: 2140-2148.
- Planet E, Attolini CS, Reina O, Flores O, Rossell D. 2012. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**: 589-590.

- Portela A, Liz J, Nogales V, Setien F, Villanueva A, Esteller M. 2013. DNA methylation determines nucleosome occupancy in the 5'-CpG islands of tumor suppressor genes. *Oncogene* **32**: 5421-5428.
- Portella G, Battistini F, Orozco M. 2013. Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. *PLoS Comput Biol* **9**: e1003354.
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.
- Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25.
- Rutledge MT, Russo M, Belton J-M, Dekker J, Broach JR. 2015. The yeast genome undergoes significant topological reorganization in quiescence. *Nucleic Acids Research* **43**: 8299–8313.
- Schlenstedt G, Hurt E, Doye V, Silver PA. 1993. Reconstitution of nuclear protein transport with semi-intact yeast cells. *J Cell Biol* **123**: 785-798.
- Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. 2017. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**: e1005665.
- Shen L, Gao G, Zhang Y, Zhang H, Ye Z, Huang S, Huang J, Kang J. 2010. A single amino acid substitution confers enhanced methylation activity of mammalian Dnmt3b on chromatin DNA. *Nucleic Acids Res* **38**: 6054-6064.
- Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**: 204-220.
- Subramaniam D, Thombre R, Dhar A, Anant S. 2014. DNA Methyltransferases: A Novel Target for Prevention and Therapy. *Frontiers in oncology* **4**: 80.
- Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465-476.
- Team RDC. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
- Varoquaux N, Ay F, Noble WS, Vert JP. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**: i26-33.
- Winkelman J, Lin L, Schormair B, Kornum BR, Faraco J, Plazzi G, Melberg A, Cornelio F, Urban AE, Pizza F et al. 2012. Mutations in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy. *Hum Mol Genet* **21**: 2205-2210.
- Zhang Z, Li G, Toh KC, Sung WK. 2013. 3D chromosome modeling with semi-definite programming and Hi-C data. *J Comput Biol* **20**: 831-846.

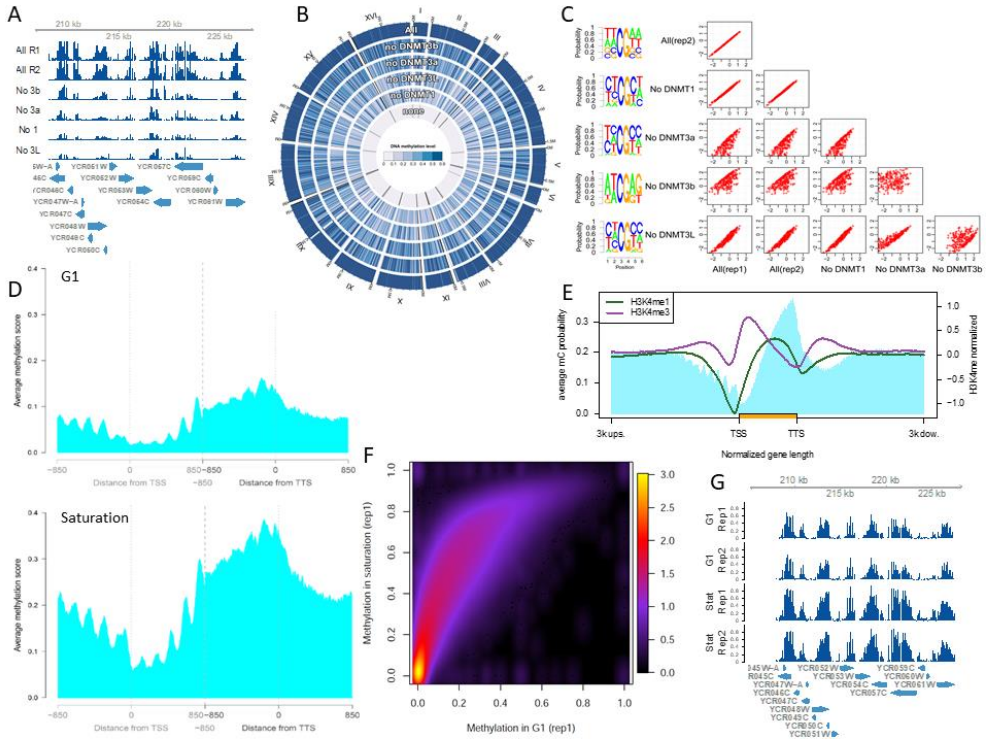


Figure 1. Methylation pattern across several samples and along the gene body. (A) The pattern of methylation is conserved in all samples as illustrated for this 20kb region of chromosome III (208135..227458) where the level of methylation at each position is represented for two samples with the four DNMTs expressed and four samples with each combination of 3 DNMTs. (B) Circular plot comparing DNA methylation for the samples in (A) and a control without methylation (None). Methylation levels decrease when one DNMT is missing, the strongest effect being without DNMT1 and the mildest effect when DNMT3b is not present. (C) Correlation between the sequence effect on methylation status among samples with different combinations of three DNMTs. Motif effects are estimated from logistic regression (for details see Materials and Methods) and correlation plots are produced for each pair of samples. Motifs have nearly the same effect in the two replicates with all DNMTs induced (correlation coefficient is cor=0.999) and when DNMT1 (cor=0.996) or DNMT3L (cor=0.954) are removed. In contrast, the estimated effect of some motifs on methylation probability is different in samples lacking DNMT3a (cor=0.793) or DNMT3b (cor=0.631). The left panels show the sequence logo of the motifs preferentially methylated in each sample. (D,F,G) Comparison of methylation pattern in samples in exponential phase and at saturation (D) Average methylation level around TSS and TTS (850 bp upstream and downstream from each point). (E) Superposition of the pattern of DNA methylation and the pattern of H3K4 methylation along the average gene body (from 3kb upstream TSS to 3kb downstream TTS). DNA methylation preferentially occurs where H3K4 is not methylated. (F) Heatmap showing the correlation between methylation probabilities in samples in G1 and at saturation. (G) The pattern of methylation is conserved in samples in G1 and samples at saturation as illustrated for this 20Kb region of chromosome III (208135..227458) where the level of methylation at each position is represented for 2 replicates of each condition.

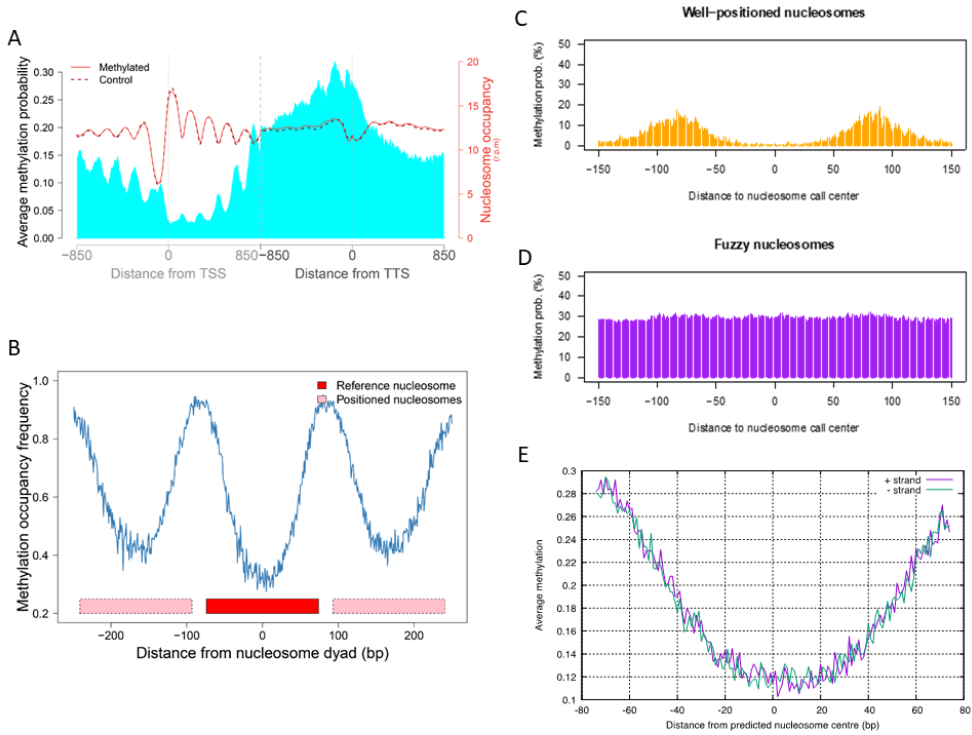


Figure 2. Correlation between DNA methylation and nucleosome coverage genome wide
 (A) Nucleosome positioning (in red) in a sample before (dashed lane) and after (plain lane) induction of methylation and average methylation probabilities (in blue). Plots are built around TSS and TTS (850 bp upstream and downstream from each point). Average nucleosome positioning does not change drastically upon methylation. (B) Percentage of CpG with methylation probability above 0.01 around well positioned nucleosomes. Nucleosome calls were considered well-positioned (W) or fuzzy (F) when nucleR peak width score and height score were higher than 0.6 and 0.4, respectively. DNA methylation is anti-correlated with nucleosome occupancy in W nucleosomes. (C,D) Average methylation probability around nucleosome call center (150 bp upstream and downstream) for (C) W and (D) F nucleosomes. (E) Average methylation probability per strand around nucleosome call center (75 bp upstream and downstream) of well-positioned nucleosomes

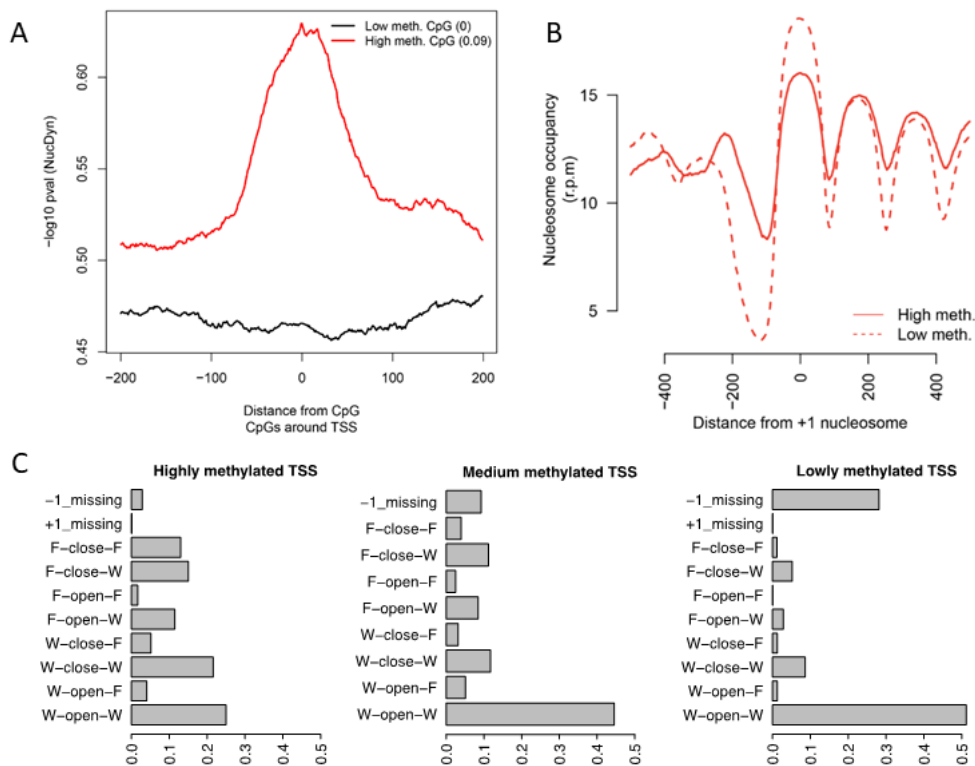


Figure 3. Nucleosome dynamics upon methylation at promoters. (A) NucDyn score around highly methylated and lowly methylated CpGs at promoters. (B) Nucleosome coverage around +1 nucleosome for genes with highly or lowly methylated promoters. (C) Nucleosome architecture around promoters according to their methylation level.

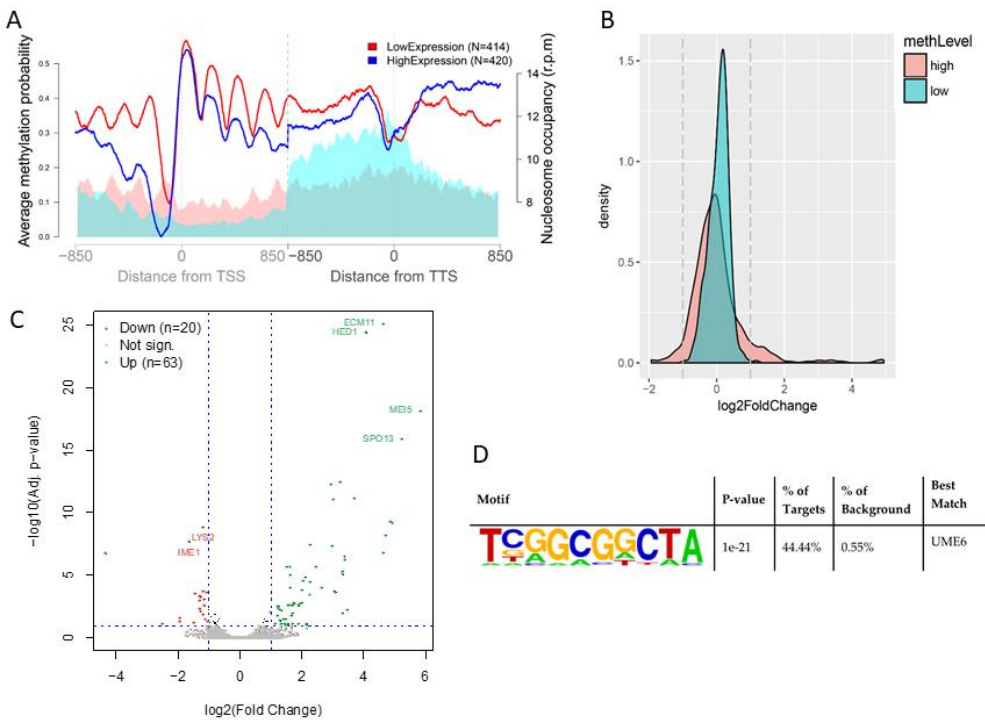


Figure 4. Correlation of DNA methylation, nucleosome positioning and gene expression. (A) Methylation probability and nucleosome coverage (solid lines) around genes with expression in the top 10% (blue) or bottom 10% (red) expression level. (B) Log2 of the fold change in expression for genes with highly and lowly methylated promoters. (C) Gene expression difference between methylated and not methylated samples. RNA level difference is plotted on the x-axis and the Adj. p-value on the y-axis. Downregulated (20 genes) and upregulated (63 genes) genes are shown in red and green, respectively. The genes with the highest changes are highlighted. (D) Promoter motif enrichment for genes with highly methylated promoters and increase in expression level.

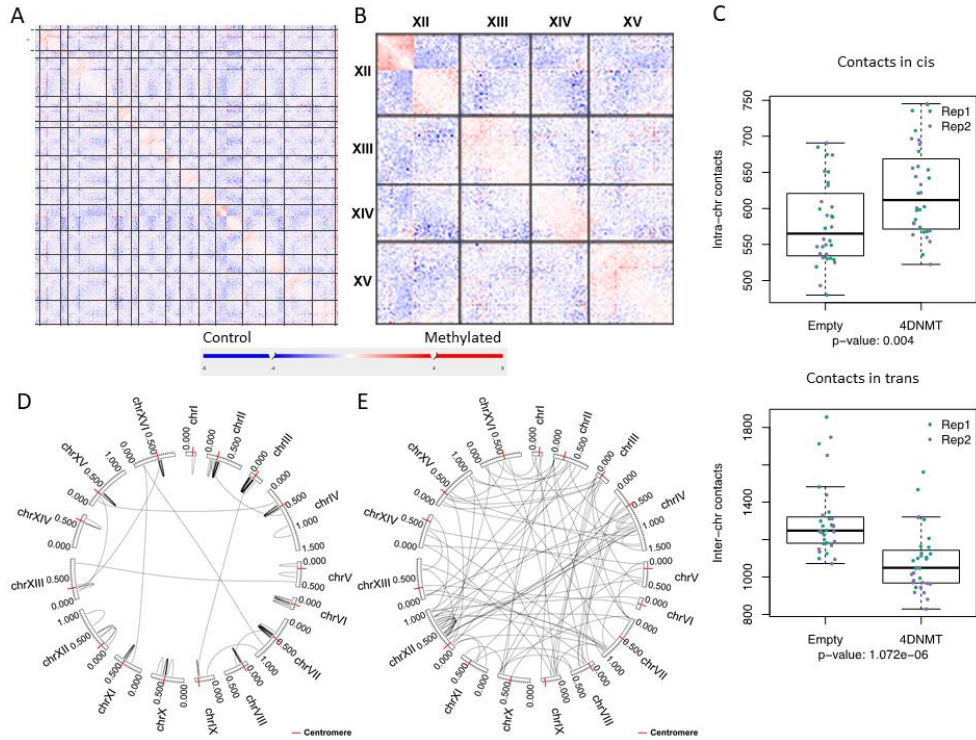


Figure 5. Effect of DNA methylation on 3D genome structure (A,B) Differential contact frequencies in control and methylation induced samples in replica 1 for (A) whole genome and (B) focus on four chromosomes. Blue indicates interaction with a higher frequency in the non-methylated control sample and red indicates interactions with a higher frequency in the methylated samples. (C) Comparison of contact frequencies between control and methylated Hi-C samples in *cis* (+/- 50Kb from the centromere, top panel) and in *trans* (lower panel). (D,E) Circos plots displaying the significant (FDR<0.5) differential interactions identified with diffHiC: (D) gained interactions ($\log_2FC > 1$) are clustered around the centromeres (red tick marks) and (E) lost interactions ($\log_2FC < -1$) preferentially occur between chromosomes.

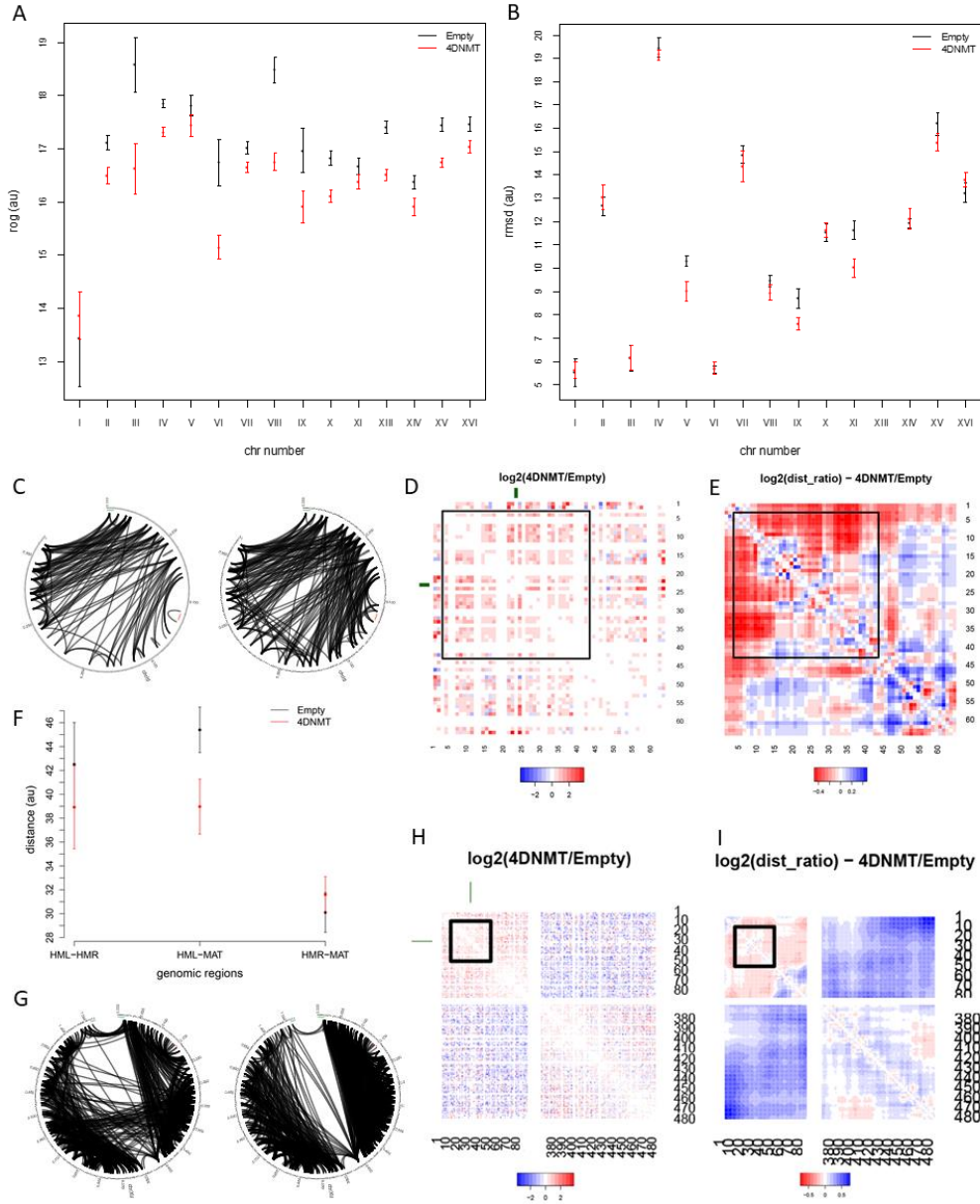


Figure 6. Chromosome conformation changes under DNA methylation. Structural changes measured on the ensemble of structures obtained with our restraint-based 3D model for each chromosome: (A) Mean radius of gyration computed around the centromeres (± 100 Kb) and (B) flexibility of each chromosome measured by the RMSD for the control (black) and methylated (red) samples. (C) Circos diagrams of significant interactions in chromosome III for the control (left) and methylated (right) samples. (D,E) Heatmaps displaying the log2 ratio (Methylated/Control) of (D) the contact frequencies and (E) the distances in the ensemble for chromosome III. Blue indicates interaction with a higher frequency or shorter distance in the non-methylated control sample and red indicates interactions with a higher frequency or shorter

distance in the methylated samples. (F) Average distances between matting type loci in the ensemble of structures for chromosome III. (G) Circos diagrams of significant interactions in chromosome XII for the control (left) and methylated (right) samples. (H,I) Log2 ratio (Methylated/Control) of (H) the contact frequencies and (I) the distances in the ensemble for chromosome XII.

Table 1: Average methylation in CpG and non-CpG context.

Sample	DNMT expressed	Hours of induction	State of the culture	All Contexts			CpG Contexts			Non-CpG Contexts		
				Avg. meth	No. Cyt	Frac. > 0	Avg. meth	No. Cyt	Frac. > 0	Avg. meth	No. Cyt	Frac. > 0
T859	DNMT1, 3a, 3L	30 hrs	Not synchronized	0.75%	3066478	2.74%	3.14%	525937	15.60%	0.16%	2538540	0.08%
T860	DNMT1, 3b, 3L	30 hrs	Not synchronized	0.70%	2584192	2.65%	2.77%	463127	14.36%	0.13%	2118519	0.10%
T861	DNMT3a, 3b, 3L	30 hrs	Not synchronized	0.49%	2889913	2.03%	1.97%	502806	11.46%	0.11%	2385079	0.04%
T869	DNMT1, 3a, 3b	30hrs	Not synchronized	0.56%	3066743	1.86%	2.08%	524612	10.73%	0.18%	2539830	0.03%
T870	None	30hrs	Not synchronized	0.15%	2732598	0.00%	0.13%	464513	0.00%	0.16%	2265439	0.00%
T862	All ¹	27.5 hrs	Not synchronized	2.03%	2810320	6.66%	8.55%	506621	34.88%	0.24%	2301426	0.45%
T863	All ²	27.5 hrs	Not synchronized	2.13%	2937375	6.90%	9.14%	522749	36.43%	0.26%	2412449	0.51%

¹Samples corresponding to replica1

²Samples corresponding to replica2

Bibliography Chapter 6

- [1] M. V. C. Greenberg and D. Bourc'his, "The diverse roles of DNA methylation in mammalian development and disease," *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 10, pp. 590–607, Oct. 2019.
- [2] A. Pérez *et al.*, "Impact of Methylation on the Physical Properties of DNA," *Biophys. J.*, vol. 102, no. 9, pp. 2140–2148, May 2012.
- [3] G. Portella, F. Battistini, and M. Orozco, "Understanding the Connection between Epigenetic DNA Methylation and Nucleosome Positioning from Computer Simulations," *PLoS Comput. Biol.*, vol. 9, no. 11, p. e1003354, Nov. 2013.
- [4] C. K. Collings and J. N. Anderson, "Links between DNA methylation and nucleosome occupancy in the human genome," *Epigenetics Chromatin*, vol. 10, no. 1, Dec. 2017.
- [5] M. M. Suzuki and A. Bird, "DNA methylation landscapes: provocative insights from epigenomics," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 465–476, Jun. 2008.
- [6] X. Yang, H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones, and G. Liang, "Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer," *Cancer Cell*, vol. 26, no. 4, pp. 577–590, Oct. 2014.
- [7] M. Morselli *et al.*, "In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse," *eLife*, vol. 4, p. e06205, Apr. 2015.
- [8] L. Lazar-Stefanita *et al.*, "Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle," *EMBO J.*, vol. 36, no. 18, pp. 2684–2697, Sep. 2017.
- [9] S. A. Schalbeter *et al.*, "SMC complexes differentially compact mitotic chromosomes according to genomic context," *Nat. Cell Biol.*, vol. 19, no. 9, pp. 1071–1080, Aug. 2017.

Chapter 7 . General discussion and conclusions

Understanding the complex mechanisms of gene regulation in the nucleus requires a detailed knowledge of chromatin structure and this implies the study of DNA at different levels of resolution, from atomistic details up to whole genome organization. In this thesis, several studies have been performed in order to analyze genome organization based on DNA intrinsic factors determined by the nucleotide sequence as well as extrinsic features such as histones, transcription factors or RNA polymerase.

7.1 Sequence dependent DNA flexibility and protein recognition

The development of a new accurate forcefield for Molecular Dynamics (MD) simulations by our group has allowed the structural analysis of trajectories of many DNA sequences that provide input for the study of sequence dependent properties of the DNA. In this thesis, three publications (Chapters 3 and 4) address the role of this intrinsic features on DNA structure, protein binding and nucleosome formation.

In the first publication, we characterized a tetra-nucleotide sequence that was previously identified to be unusually flexible and for which it was not possible to understand its dynamics using available dimer or tetramer models. We analyzed the structural polymorphisms of this tetramer in different sequence contexts, considering long range (beyond the tetramer level) sequence effects by means of MD

simulations, as well as from data mining of experimental structures deposited in PDB. The flexibility inherent to this tetramer implies that it can be present in the chromatin in very different states and this might have impact in genome structure which should be reflected in its prevalence. We observed that this tetramer is rather infrequently found in the genome of several eukaryotes, despite containing one of the stop codons, it is enriched in intergenic regions and depleted in coding sequences, and it has low mutation rate in different cancer types compared to other tetramers. Our results suggest that its unique conformational properties might be important for its significant underrepresentation in the genome.

The second publication shows that the sequence dependent structural flexibility is also important for protein recognition of target binding sites. Consensus sequences for a large number of proteins have been identified, but the mechanism of recognition is not well understood. Here, using the physical properties of DNA and theoretical studies based on MD simulations we have found prevalence of conformational selection in many protein-DNA complexes from structures in the PDB. This implies that most of the motifs can spontaneously sample the conformation required for protein binding, reducing the prevalence of the induce-fit paradigm to a minority of cases, where specific backbone rearrangements are required leading to strong disruptions of the DNA structure.

Finally, in the third publication we have used the physical descriptors obtained from the MD simulations to study the deformation energy of the DNA in the nucleosome, that is a key element to understand most of the processes in the nucleus required for cell functioning. We demonstrated the existence of energetic barriers that define the positioning of the two nucleosomes at the 5' (+1 nucleosome) and 3' (-last nucleosome) gene ends in the yeast genome. Although previous studies obtained low accuracy predicting nucleosome organization from the sequence dependent features, our study shows that combined with protein binding affinity scores we could predict with good accuracy the position of nucleosome free regions (NFR) at the transcription start site and transcription termination site. These two barriers define the position of the +1 and -last nucleosome in the gene, for which the nucleosome organization along the gene body can be predicted by signal theory using two

periodic signals running in opposite direction from the +1 and -last nucleosomes. When the two signals are in phase, the nucleosomes are well-positioned along the gene body. On the contrary, anti-phased signals produce fuzzier configurations. A series of synthetic biology experiments, followed by computational analysis of the obtained profiles, showed that altering the periodicity does not lead to differential expression, but gene regulation is more determinant on nucleosome positioning. We also demonstrated that ordered nucleosome string in the gene body correlates with active genes. A series of experiments complemented with bioinformatics analysis uncover the causal relationship: more polymerase activity → higher nucleosome ordering.

7.2 Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning

Besides theoretical study of nucleosome positioning, along this thesis we have analyzed several MNase-seq experiments performed under different conditions. In our group, a software for the mapping of nucleosome positions from this experimental technique, nucleR, was developed several years ago. Although it allows to study the nucleosome organization in an experiment very accurately, it cannot perform direct comparison between two different experimental conditions. Moreover, since the MNase-seq data come from a population of cells, the noise sometimes masks the real differences occurring between two experimental conditions and adding up the coverage for all cells obstructs the detection of changes. For this reason, we developed a new algorithm, NucDyn, that works at the fragment level, to capture variability from the different cells in the experiment. Comparing our results with other software on synthetically produced nucleosome maps, we found that NucDyn is superior to detect nucleosome rearrangements affecting a part of the cell population.

Nucleosome Dynamics package, comprising NucDyn together with nucleR and other tools that we developed for the analysis of nucleosome positioning (e.g. classification of the NFR around the TSS, nucleosome periodicity, nucleosome

stiffness), has been integrated into a virtual research environment (MuGVRE). This framework allows not only easy and automatized analyses of nucleosome experimental data, but also to put the results in the context of genomic information (ChIP-seq, DNA methylation, etc.) relevant to understand the role of nucleosome organization in different cellular processes. For instance, analyses performed with our package for MNase-seq experiments on different stages of the cell cycle, along the yeast metabolic cycle or in different sources of carbon showed important nucleosome rearrangements in promoters of genes that are activated or repressed in response to the different conditions.

7.3 Impact of DNA methylation on 3D genome structure

DNA methylation can influence chromatin organization and DNA. Previous *in vitro* and *in silico* studies found increasing DNA stiffness due to CpG methylation at the local level. We were interested in understanding how it might affect chromatin structure at larger scale: at the nucleosome level and whole genome 3D structure. To perform this analysis, we used an organism that is natively unmethylated, *Saccharomyces cerevisiae*, and induced DNA methylation expressing four DNMTs, allowing us to directly study the effect of this epigenetic factor on chromatin, removing the effect of methylation readers present in more complex organisms.

Although yeast does not have any of the machinery required to read or write the DNA methylation fingerprint, the pattern observed along genes is similar in other organisms that have the DNA methylation machinery. This shows that histone marks such as H3K4 methylation could be important in writing the DNA methylation at the correct positions by direct effects which should be related to the different binding affinities of normal and methylated DNAs. Our results suggest that although DNA methylation can alter the physical properties of DNA producing more fuzzy nucleosome profiles, the global pattern of nucleosome occupancy is not largely altered, which explains cell viability. However, for the promoters with largest levels of DNA methylation, we identified large changes in nucleosome positioning. Several genes are repressed upon methylation, which can be explained considering the steric hindrance that a displaced nucleosome generates, but those which are activated are

more difficult to understand. We found that these genes share a common motif (URS1, that is a binding site for UME6 protein which represses the expression of those genes) containing CpG steps that are highly methylated. We hypothesized that the differential expression could be due to the inability to bind to the target motif due to the methylation and therefore the genes cannot be repressed. This hypothesis is supported by the fact that the expression level is highly correlated with the methylation level at those sites. Again, this relationship can be explained only by the different protein-binding properties of normal and methylated DNA as no methylated-recognition protein exist in yeast.

Then, we studied at the large scale the 3D conformation of chromatin changes using Hi-C data. Upon methylation less inter-chromosome contacts are observed, and chromosomes become more condensed, especially around the centromere. An exception is chromosome XII, containing the rDNA region, that forms a barrier separating the two ends of the chromosome in the methylated sample, but allows the contacts between the two regions in the control sample in saturation. We built a restraint-based model from the contact matrices. It confirmed the differential structure around the centromeric regions, showing decrease in the radius of gyration, and the segregation of the two regions separated by the rDNA in chromosome XII. Another chromosome where many significant differential interactions are observed is chromosome III. Interestingly, it contains the heterochromatic regions of matting type loci. Moreover, the telomere-telomere contacts are also reduced in the presence of DNA methylation. These results suggest that the chromatin structure is blocked in heterochromatic regions upon methylation while the cells are dividing, keeping the heterochromatic regions segregated after the cells enter the stationary phase.

In summary, our analysis revealed the intrinsic effect of DNA methylation on chromatin organization, independent of the effect of DNA methylation readers that recognize methylation signatures, which are absent in our model organism.

Conclusions

- The study of the unusually flexible CTAG tetramer reveals that its unique conformational properties might have impact in genome structure, reflected in its significant underrepresentation in the genome.
- The conformational selection protein readout mechanism is prevalent in the recognition of DNA by effector proteins, except in a few specific cases where base opening or extreme distortions of the fiber are required.
- A machine learning algorithm was proposed for the detection of nucleosome free regions, based on the deformation energy of the DNA and transcription factor binding affinity. It performs accurately in the yeast genome and allows to identify barriers from which periodic signals are sent to define the nucleosome architecture at gene bodies.
- NucDyn, an algorithm for the detection of changes in nucleosome architecture comparing two MNase-seq experiments was developed. It can find differences occurring even in small percentages of the cells, outperforming other available methods. This algorithm and other tools for the analysis of nucleosome positioning have been integrated into a package called Nucleosome Dynamics, available through different distribution models (R packages, web-servers, containerized distributions). In particular, the implementation in the MuGVRE showed to be useful for the analysis of three test cases where the changes were correlated to response to different cell conditions.
- We explored the intrinsic effect of DNA methylation on nucleosome positioning using an engineered yeast to which we transferred all methylation machinery. Although it is not a common universal reorganization, increase in fuzziness is observed and at some specific promoters the high methylation and nucleosome displacements are related to changes in gene expression. The 3D chromatin model developed based on restraints from Hi-C experiments allows to obtain a set of structures that represent a high percentage of the observed experimental contacts. With this model, we observed that methylation induces the

reorganization of chromatin at the intra and inter chromosomal levels. Globally, more contacts are formed around the centromeres while inter-chromosomal contacts are reduced. Moreover, we found that methylation is important in maintaining the structure in heterochromatin regions in chromosomes III and XII and at telomeric regions.

Resumen

Comprender la conexión entre la organización del ADN en el núcleo y el funcionamiento celular es uno de los problemas más interesantes en biología. Aunque se han desarrollado muchos esfuerzos interdisciplinarios para este objetivo, los mecanismos de plegamiento del ADN son en gran medida desconocidos. Por lo tanto, la complejidad de la estructura del genoma requiere diferentes técnicas para abordar varios niveles de resolución.

En esta tesis, se estudian varias escalas de plegamiento del genoma utilizando métodos teóricos. Primero, nos centramos en las propiedades dependientes de la secuencia de ADN que definen la propensión de regiones específicas a ser reconocidos por las proteínas, descubriendo que la flexibilidad de ciertas secuencias de ADN podría explicar su prevalencia en el genoma.

Las propiedades dependientes de la secuencia de ADN también son importantes para definir la primera capa de organización de la cromatina: el nucleosoma. Los descriptores físicos de la secuencia de ADN combinados con la propensión a la unión de factores de transcripción son muy informativos sobre la posición de las regiones no afines a la formación de nucleosomas, que guían la posición de los nucleosomas +1 y -último, y el resto de los nucleosomas en el cuerpo del gen se coloca por posicionamiento estadístico. Existe una clara correlación entre la actividad transcripcional y la fase de nucleosomas en el cuerpo del gen, encontrando que la transcripción influye más sobre la organización de los nucleosomas que la relación opuesta.

En esta tesis también se desarrolló un paquete para el análisis comparativo de la organización de nucleosomas que permite predecir cuantitativamente

los cambios en el posicionamiento de los nucleosomas que ocurren cuando se introducen perturbaciones en la célula.

Finalmente, estudiamos tanto los cambios a nivel de nucleosomas como a mayor escala producidos por la inducción de la metilación del ADN en un genoma que originalmente no tiene metilación, desarrollando un modelo 3D basado en Hi-C para estudiar la reorganización de la cromatina. Encontramos cambios muy significativos en la estructura de la cromatina inducidos por la metilación, que se reflejan en la expresión génica y el fenotipo celular. Curiosamente, estos cambios se encuentran en un organismo modelo que no tiene proteínas preparadas para reconocer la metilación y, en consecuencia, pueden deberse a los efectos intrínsecos (no mediados por proteínas) de la metilación.

Introducción

El ADN es una molécula larga que, en condiciones fisiológicas, forma un dúplex complementario que contiene la información genética necesaria para construir la vida. Aunque la fibra de ADN humano tiene aproximadamente dos metros de largo, está compactada para ajustarse dentro del pequeño espacio definido por el núcleo celular con un diámetro de aproximadamente 10 micrómetros [1]. La compactación del ADN es mediada por proteínas que guían su plegamiento dentro del núcleo de las células eucariotas. El complejo de ADN y proteínas dentro del núcleo se conoce como cromatina. Muchas evidencias experimentales [2]–[4] demuestran que el empaquetamiento del ADN dentro del núcleo no es aleatorio, ya que se debe preservar la accesibilidad al ADN a los reguladores del genoma, asegurando la función correcta de procesos como la transcripción, la replicación y la reparación del ADN. Otras evidencias han demostrado que esta organización es dinámica y sufre diferentes reorganizaciones a lo largo de varios procesos celulares como la diferenciación [2], la progresión del ciclo celular [5] o la respuesta al daño celular [6].

La unidad fundamental de compactación del ADN en organismos eucariotas es el nucleosoma. Un nucleosoma canónico está formado por ~ 147 pares de bases (bp) de ADN bicatenario que se enrollan en aproximadamente 1.65 vueltas súper helicoidales alrededor de dos copias de cada histona H2A, H2B, H3 y H4. La curvatura del ADN en el nucleosoma requiere una energía de flexión significativa [7].

Las posiciones de los nucleosomas *in vivo* se han determinado utilizando varios protocolos experimentales, como FAIRE [8], ATAC-seq [9] y MNase-seq [10]. Esta última es la técnica más utilizada y proporciona información detallada sobre la organización de los nucleosomas. Estos experimentos

contienen información de una población de células, por lo tanto, los perfiles de nucleosomas pueden ser ruidosos [11] y se caracterizan típicamente por dos propiedades importantes: ocupación y posicionamiento. El primero está relacionado con el porcentaje de células en un experimento que contiene un nucleosoma dado, el último denota la variabilidad en su posición genómica entre todas las células. Un nucleosoma se llama bien posicionado (W) cuando está presente en un gran porcentaje de las células, y los fragmentos de diferentes células presentan baja variabilidad con respecto a la posición genómica. Cuando un nucleosoma tiene baja cobertura y / o gran variabilidad de posicionamiento, se llama difuso (F) [11].

La organización de los nucleosomas en la secuencia de ADN no es aleatoria y se ha relacionado con diferentes procesos celulares como la transcripción y la replicación [12]. Además, es dinámico en el espacio y el tiempo, y está influenciado por varios factores, tales como: (i) el contexto local determinado por propiedades dependientes de la secuencia (factores *cis*), (ii) complejos de proteínas que interactúan con el ADN y pueden competir con nucleosomas (factores *trans*), como factores de transcripción [13], maquinaria de replicación [12] o remodeladores dependientes de ATP que pueden deslizar o expulsar nucleosomas (parcial o totalmente) [14], y (iii) el efecto de los nucleosomas vecinos que imponen restricciones estéricas para el posicionamiento de nucleosomas [15].

A escala global los cromosomas se pliegan jerárquicamente en el espacio nuclear durante la interfase [16], [17]. A nivel de todo el núcleo, la cromatina está segregación en territorios cromosómicos [18]. A mayor escala, se ha observado la separación de los compartimentos A / B, que corresponden a la eucromatina transcrita activamente y la heterocromatina reprimida, respectivamente [19], estando esta última unida preferentemente a la lámina nuclear [20].

A una escala más fina, los cromosomas se organizan en dominios asociados topológicamente (TAD), regiones del genoma con alta auto-interacción, aisladas de regiones de dominios vecinos [21]. Los TADs podrían reflejar la presencia de bucles de genes que imponen la direccionalidad del promotor [22]o bucles formados para atraer elementos reguladores en proximidad de los genes sobre los que influyen, que pueden estar separados por una gran distancia genómica, [23], [24]. Los bordes de los TADs en células de mamíferos se colocalizan fuertemente con los sitios de unión de CTCF [25]. Un mecanismo propuesto para la formación de TAD implica el papel de los factores de extrusión, como la cohesina, que extruyen el ADN a través de su estructura en forma de anillo [26], [27]. La formación del bucle continúa hasta que el factor de extrusión encuentra otro factor límite, por ejemplo, CTCF en orientación convergente en los bordes del TAD.

Objetivos

El objetivo principal de esta tesis es estudiar la estructura y organización de la fibra de ADN a diferentes niveles de detalle, desde propiedades específicas de secuencia local hasta la estructura 3D global dentro del núcleo. Para este propósito, los siguientes objetivos específicos se proponen y agrupan en tres categorías:

1. Propiedades dependientes de la secuencia de ADN
 - Caracterizar la distribución amplia del genoma y la función de secuencias de ADN altamente flexibles.
 - Evaluar los mecanismos para el reconocimiento de ADN de proteínas que definen pruebas estadísticas para la detección de diferencias significativas en los descriptores físicos de ADN entre la estructura experimental de ADN unida a proteínas y la estructura desnuda a partir de simulaciones de dinámica molecular.
 - Para predecir perfiles de organización de nucleosomas utilizando métodos de aprendizaje automático basados en la energía de deformación del ADN, la afinidad del factor de transcripción y la periodicidad de la señal de nucleosoma.
2. Herramientas para estudiar el posicionamiento de nucleosomas in vivo
 - Desarrollar un algoritmo para comparar perfiles de posicionamiento de nucleosomas entre dos poblaciones celulares.
 - Integrar diferentes herramientas para el análisis de la organización de nucleosomas en una tubería disponible a

través de diferentes modelos de distribución (servidores web, distribuciones en contenedores) que facilitan el análisis de resultados en el contexto de otra información genómica.

3. Efecto de la metilación del ADN en la estructura de la cromatina.

- Analizar el efecto de la metilación del ADN en el posicionamiento de nucleosomas in vivo, aplicando el algoritmo propuesto para la comparación de perfiles de nucleosomas.
- Estudiar los cambios de cromatina a nivel de estructura 3D del genoma completo aplicando métodos estadísticos para la detección de regiones de interacción diferencial en datos de Hi-C.
- Desarrollar un modelo 3D de grano grueso de la cromatina basado en restricciones obtenidas de matrices de contacto Hi-C para un análisis posterior de los cambios estructurales producidos por la metilación del ADN.

Discusión general

La comprensión de los complejos mecanismos de regulación génica en el núcleo requiere un conocimiento detallado de la estructura de la cromatina y esto implica el estudio del ADN a diferentes niveles de resolución, desde detalles atomísticos hasta la organización del genoma completo. En esta tesis, se han realizado varios estudios para analizar la organización del genoma teniendo en cuenta tanto factores intrínsecos de ADN determinados por la secuencia de nucleótidos, así como características extrínsecas como histonas, factores de transcripción o ARN polimerasa.

Flexibilidad del ADN y reconocimiento de proteínas asociado a la secuencia

El desarrollo de un nuevo campo de fuerza para simulaciones de Dinámica Molecular (MD) por parte de nuestro grupo ha permitido el análisis estructural de las trayectorias de muchas secuencias de ADN que proporcionan información para el estudio de las propiedades dependientes de la secuencia del ADN. En esta tesis, tres publicaciones (capítulos 3 y 4) abordan el papel de estas características intrínsecas en la estructura del ADN, la unión a proteínas y la formación de nucleosomas.

En la primera publicación, caracterizamos una secuencia de tetra-nucleótidos que previamente se identificó como inusualmente flexible y para la cual no fue posible comprender su dinámica utilizando los modelos de dímero o tetrámero disponibles. Analizamos los polimorfismos estructurales de este tetrámero en diferentes contextos de secuencia, considerando los efectos de secuencia de largo alcance (más allá del nivel del tetrámero) por medio de simulaciones MD, así como de la minería de datos de estructuras

experimentales depositadas en la base de datos *Protein Data Bank* (PDB). La flexibilidad inherente a este tetrámero implica que puede estar presente en la cromatina en estados muy diferentes y esto podría tener un impacto en la estructura del genoma que debería reflejarse en su prevalencia. Observamos que este tetrámero es poco frecuente en el genoma de varios organismos eucariotas, a pesar de contener uno de los codones de parada (TAG), está enriquecido en regiones intergénicas y empobrecido en secuencias codificantes, y tiene una baja tasa de mutación en diferentes tipos de cáncer en comparación con otros tetrámeros. Nuestros resultados sugieren que sus propiedades conformacionales únicas podrían ser importantes para su significativamente baja representación en el genoma.

La segunda publicación muestra que la flexibilidad estructural dependiente de la secuencia también es importante para el reconocimiento de proteínas de los sitios de unión al ADN. Se han identificado secuencias de consenso para un gran número de proteínas, pero el mecanismo de reconocimiento no ha sido establecido. En este trabajo, utilizando las propiedades físicas del ADN y estudios teóricos basados en simulaciones MD, hemos encontrado la prevalencia de la selección conformacional en muchos complejos de proteína-ADN de estructuras en el PDB. Esto implica que la mayoría de los motivos pueden muestrear espontáneamente la conformación requerida para la unión a proteínas, reduciendo la prevalencia del paradigma de ajuste inducido a una minoría de casos, donde se requieren reorganizaciones específicas del esqueleto que conducen a importantes modificaciones de la estructura del ADN.

Finalmente, en la tercera publicación, hemos utilizado los descriptores físicos obtenidos de las simulaciones de MD para estudiar la energía de deformación del ADN en el nucleosoma, que es un elemento clave para comprender la mayoría de los procesos en el núcleo necesarios para el funcionamiento

celular. Demostramos la existencia de barreras energéticas que definen el posicionamiento de los dos nucleosomas en los extremos 5' (+1 nucleosoma) y 3' (-último nucleosoma) de cada gen en el genoma de la levadura (*Saccharomyces cerevisiae*). Aunque estudios anteriores obtuvieron predicciones con poca precisión de la organización de nucleosomas a partir de las propiedades físicas del ADN, nuestro estudio muestra que, combinados con medidas de afinidad de la unión a proteínas, podemos predecir con buena precisión la posición de las regiones libres de nucleosomas en el sitio de inicio de la transcripción y el sitio de terminación de la transcripción. Estas dos barreras definen la posición del nucleosoma +1 y -último en el gen, con lo cual es posible predecir la organización de nucleosomas a lo largo del cuerpo del gen mediante la teoría de la señal utilizando dos señales periódicas que se envían en dirección opuesta a partir de los nucleosomas +1 y -último. Cuando las dos señales están en fase, los nucleosomas están bien posicionados a lo largo del cuerpo del gen. Por el contrario, las señales en anti-fase producen configuraciones de nucleosomas más difusas. Una serie de experimentos de biología sintética, seguida de un análisis computacional de los perfiles obtenidos, mostró que alterar la periodicidad no conduce a la expresión diferencial, pero la regulación génica es más determinante en el posicionamiento de nucleosomas. También demostramos que la cadena de nucleosomas ordenada en el cuerpo del gen se correlaciona con genes activos. Una serie de experimentos complementados con análisis bioinformáticos descubren la relación causal: más actividad de polimerasa mayor ordenación de nucleosomas.

Nucleosome Dynamics: una nueva herramienta para el análisis dinámico del posicionamiento de nucleosomas

Además del estudio teórico del posicionamiento de nucleosomas, a lo largo de esta tesis se han analizado varios experimentos de MNase-seq realizados en diferentes condiciones. En nuestro grupo, hace varios años se desarrolló un software para el mapeo de las posiciones de nucleosomas a partir de datos obtenidos mediante esta técnica experimental, nucleR. Aunque permite estudiar la organización de nucleosomas en un experimento con mucha precisión, no puede realizar una comparación directa entre dos condiciones experimentales diferentes. Además, dado que los datos de MNase-seq provienen de una población de células, el ruido a veces oculta las diferencias reales que ocurren entre dos condiciones experimentales y sumar la cobertura para todas las células impide la detección de cambios. Por esta razón, desarrollamos un nuevo algoritmo, NucDyn, que funciona a nivel de los fragmentos secuenciados, para capturar la variabilidad de las diferentes células en el experimento. Al comparar nuestros resultados con otros programas computacionales en mapas de nucleosomas producidos sintéticamente, encontramos que NucDyn es superior para detectar reordenamientos de nucleosomas que afectan a una parte de la población celular.

El paquete Nucleosome Dynamics, que comprende NucDyn junto con nucleR y otras herramientas que desarrollamos para el análisis del posicionamiento de nucleosomas (por ejemplo, clasificación de las regiones libres de nucleosomas alrededor de los inicios de transcripción, computo de medidas de periodicidad de nucleosomas, estimación de la rigidez asociada a los nucleosomas), se ha integrado en un entorno de investigación virtual (MuGVRE). Esta web permite no solo realizar el análisis de forma automática de datos experimentales de nucleosomas, sino también poner los resultados en el contexto de información genómica (ChIP-seq, metilación de ADN, etc.) lo cual es relevante para comprender el papel de la organización de nucleosomas en diferentes procesos celulares. Por ejemplo, los análisis

realizados con Nucleosome Dynamics para experimentos MNase-seq en diferentes etapas del ciclo celular, a lo largo del ciclo metabólico de la levadura o en diferentes fuentes de carbono mostraron importantes reordenamientos de nucleosomas en promotores de genes que se activan o reprimen en respuesta a las diferentes condiciones experimentales.

Impacto de la metilación del ADN en la estructura 3D del genoma

La metilación del ADN puede influir en la organización de la cromatina y el ADN. Estudios previos *in vitro* e *in silico* encontraron un aumento de la rigidez del ADN debido a la metilación de los pasos CpG a nivel local. Sin embargo, el efecto de la metilación en la estructura global de la cromatina no ha sido establecido. Por esta razón, en esta tesis se busca comprender cómo afecta la estructura de la cromatina a mayor escala: a nivel de nucleosomas y la estructura 3D del genoma completo. Para realizar este análisis, se utilizó un organismo que no contiene los factores necesarios para producir la metilación del ADN, *Saccharomyces cerevisiae*, y se indujo expresando cuatro metiltransferasas (DNMTs), lo que nos permite estudiar directamente el efecto de este factor epigenético en la cromatina, eliminando el efecto de proteínas que reconocen la metilación, presentes en organismos más complejos.

Aunque la levadura no tiene la maquinaria necesaria para leer o escribir la metilación del ADN, el patrón observado a lo largo de los genes es similar en otros organismos que sí la tienen. Esto muestra que las marcas de histonas, como la metilación de H3K4, podrían ser importantes para escribir la metilación del ADN en las posiciones correctas por efectos directos que deberían estar relacionados con las diferentes afinidades de unión del ADN

normal y metilado. Nuestros resultados sugieren que, aunque la metilación del ADN puede alterar las propiedades físicas del ADN produciendo perfiles de nucleosomas más difusos, el patrón global de ocupación de nucleosomas no se altera en gran medida, lo que explica la viabilidad celular. Sin embargo, para los promotores con mayores niveles de metilación del ADN, identificamos grandes cambios en el posicionamiento de nucleosomas. La metilación reprime varios genes, lo cual puede explicarse considerando el obstáculo estérico que genera un nucleosoma desplazado, pero la activación de genes producida por la metilación es más difícil de entender. Analizando las funciones de estos genes, se encontró que comparten un motivo común (URS1, que es un sitio de unión para la proteína UME6 que reprime su expresión) que contiene pasos de CpG que están altamente metilados. La expresión diferencial podría deberse a la incapacidad de la proteína de reconocer y unirse a la secuencia debido a la metilación y, por lo tanto, los genes no pueden ser reprimidos. Esta hipótesis está respaldada por el hecho de que el nivel de expresión está altamente correlacionado con el nivel de metilación en esos sitios. Nuevamente, esta relación solo puede explicarse por las diferentes propiedades de unión a proteínas del ADN normal y metilado, ya que no existe proteína de reconocimiento metilado en la levadura.

Luego, estudiamos a gran escala la conformación 3D de los cambios de cromatina utilizando datos Hi-C. Tras la metilación, se observan menos contactos entre cromosomas y estos se condensan más, especialmente alrededor del centrómero. La única excepción es el cromosoma XII, que contiene las regiones de ADN ribosomal, las cuales forman una barrera que separa los dos extremos del cromosoma en la muestra metilada, pero permite la formación de contactos entre las dos regiones en la muestra de control sin metilación en saturación. Para obtener más información sobre el efecto de la metilación del ADN en la estructura de la cromatina, construimos por

primera vez un modelo basado en la restricción a partir de las matrices de contacto y los experimentos MNase-seq. Se observó diferencias en la estructura alrededor de las regiones centroméricas, mostrando una disminución en el radio de giro y la segregación de las dos regiones separadas por el ADN ribosomal en el cromosoma XII.

Conclusiones

- El estudio del tetrámero inusualmente flexible CTAG revela que sus propiedades conformacionales únicas podrían tener un impacto en la estructura del genoma, lo que se refleja en su importante subrepresentación en el genoma.
- El mecanismo de lectura de las proteínas a través de la selección conformacional prevalece en el reconocimiento del ADN por las proteínas efectoras, excepto en algunos casos específicos donde se requiere una apertura de las bases o distorsiones extremas de la fibra.
- Se propuso un algoritmo de aprendizaje automático para la detección de regiones libres de nucleosomas, basado en la energía de deformación del ADN y la afinidad de unión a factores de transcripción. Este algoritmo permite obtener predicciones con alta precisión en el genoma de la levadura e identificar barreras desde las cuales se envían señales periódicas para definir la arquitectura de nucleosomas en el cuerpo de los genes.
- Se desarrolló un método para la detección de cambios en la arquitectura de nucleosomas, NucDyn, que compara dos experimentos de MNase-seq. Puede encontrar diferencias que ocurren incluso en pequeños porcentajes de las células, superando a otros métodos disponibles. Este algoritmo y otras herramientas para el análisis de posicionamiento de nucleosomas se han integrado en un paquete llamado Nucleosome Dynamics, disponible a través de diferentes modelos de distribución (paquetes R, servidores web, distribuciones en contenedores). En particular, la implementación en MuGVRE demostró ser útil para el análisis de tres ejemplos de uso

donde los cambios se correlacionaron con la respuesta a diferentes condiciones celulares.

- Se exploró el efecto intrínseco de la metilación del ADN en el posicionamiento de nucleosomas utilizando una cepa levadura modificada a la que se transfirió la maquinaria de metilación. Aunque no se observa una reorganización universal común, se evidencia disminución en el número de nucleosomas bien posicionados. Así mismo, en algunos promotores específicos la alta metilación y los desplazamientos de nucleosomas están relacionados con cambios en la expresión génica.
- El modelo tridimensional de la cromatina desarrollado con base en restricciones obtenidas a partir de los experimentos de Hi-C permite obtener un conjunto de estructuras que representan un alto porcentaje de los contactos experimentales observados. Con este observamos que la metilación produce reorganización de la cromatina a nivel intra e inter cromosómico. Alrededor de los centrómeros se forman más contactos mientras se pierden los contactos inter cromosómicos. Adicionalmente observamos que la metilación es importante en el mantenimiento de la estructura en regiones de heterocromatina.

Bibliografía

- [1] B. R. Lajoie, J. Dekker, and N. Kaplan, "The Hitchhiker's guide to Hi-C analysis: Practical guidelines," *Methods*, vol. 72, pp. 65–75, Jan. 2015.
- [2] G. Cavalli and T. Misteli, "Functional implications of genome topology," *Nat. Struct. Mol. Biol.*, vol. 20, no. 3, pp. 290–299, Mar. 2013.
- [3] B. van Steensel and E. E. M. Furlong, "The role of transcription in shaping the spatial organization of the genome," *Nat. Rev. Mol. Cell Biol.*, Mar. 2019.
- [4] M. J. Rowley and V. G. Corces, "Organizational principles of 3D genome architecture," *Nat. Rev. Genet.*, vol. 19, no. 12, pp. 789–800, Dec. 2018.
- [5] L. Lazar-Stefanita *et al.*, "Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle," *EMBO J.*, vol. 36, no. 18, pp. 2684–2697, Sep. 2017.
- [6] E. Vizcaya-Molina, C. C. Klein, F. Serras, and M. Corominas, "Chromatin dynamics in regeneration epithelia: Lessons from *Drosophila* imaginal discs," *Semin. Cell Dev. Biol.*, p. S1084952118301952, May 2019.
- [7] R. V. Chereji and D. J. Clark, "Major Determinants of Nucleosome Positioning," *Biophys. J.*, Apr. 2018.
- [8] P. G. Giresi, J. Kim, R. M. McDaniel, V. R. Iyer, and J. D. Lieb, "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin," *Genome Res.*, vol. 17, no. 6, pp. 877–885, Jun. 2007.
- [9] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nat. Methods*, vol. 10, no. 12, pp. 1213–1218, Dec. 2013.
- [10] D. E. Schones *et al.*, "Dynamic Regulation of Nucleosome Positioning in the Human Genome," *Cell*, vol. 132, no. 5, pp. 887–898, Mar. 2008.
- [11] O. Flores, O. Deniz, M. Soler-Lopez, and M. Orozco, "Fuzziness and noise in nucleosomal architecture," *Nucleic Acids Res.*, vol. 42, no. 8, pp. 4934–4946, Apr. 2014.
- [12] W. K. M. Lai and B. F. Pugh, "Understanding nucleosome dynamics and their links to gene expression and DNA replication," *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 9, pp. 548–562, May 2017.
- [13] F. Zhu *et al.*, "The interaction landscape between transcription factors and the nucleosome," *Nature*, Sep. 2018.
- [14] G. Längst and L. Manlyte, "Chromatin Remodelers: From Function to Dysfunction," *Genes*, vol. 6, no. 2, pp. 299–324, Jun. 2015.
- [15] A. Jansen and K. J. Verstrepen, "Nucleosome Positioning in *Saccharomyces cerevisiae*," *Microbiol. Mol. Biol. Rev.*, vol. 75, no. 2, pp. 301–320, Jun. 2011.

- [16]J. H. Gibcus and J. Dekker, "The Hierarchy of the 3D Genome," *Mol. Cell*, vol. 49, no. 5, pp. 773–782, Mar. 2013.
- [17]J. Fraser *et al.*, "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation," *Mol. Syst. Biol.*, vol. 11, no. 12, pp. 852–852, Dec. 2015.
- [18]T. Cremer and C. Cremer, "Chromosome territories, nuclear architecture and gene regulation in mammalian cells," *Nat. Rev. Genet.*, vol. 2, no. 4, pp. 292–301, Apr. 2001.
- [19]E. Lieberman-Aiden *et al.*, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009.
- [20]B. van Steensel and A. S. Belmont, "Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression," *Cell*, vol. 169, no. 5, pp. 780–791, May 2017.
- [21]E. P. Nora *et al.*, "Spatial partitioning of the regulatory landscape of the X-inactivation centre," *Nature*, vol. 485, no. 7398, pp. 381–385, Apr. 2012.
- [22]S. M. Tan-Wong *et al.*, "Gene Loops Enhance Transcriptional Directionality," *Science*, vol. 338, no. 6107, pp. 671–675, Nov. 2012.
- [23]T. Sexton *et al.*, "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome," *Cell*, vol. 148, no. 3, pp. 458–472, Feb. 2012.
- [24]R. Pueschel, F. Coraggio, and P. Meister, "From single genes to entire genomes: the search for a function of nuclear organization," *Development*, vol. 143, no. 6, pp. 910–923, Mar. 2016.
- [25]S. S. P. Rao *et al.*, "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014.
- [26]G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, "Formation of Chromosomal Domains by Loop Extrusion," *Cell Rep.*, vol. 15, no. 9, pp. 2038–2049, May 2016.
- [27]A. L. Sanborn *et al.*, "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes," *Proc. Natl. Acad. Sci.*, vol. 112, no. 47, pp. E6456–E6465, Nov. 2015.

Annexes

I. Modulation of the helical properties of DNA: next-to-nearest neighbor effects and beyond

SUPPORTING INFORMATION

MODULATION OF THE HELICAL PROPERTIES OF DNA: NEXT-TO-NEAREST NEIGHBOUR EFFECTS AND BEYOND

Alexandra Balaceanu¹, Diana Buitrago¹, Jürgen Walther¹, Adam Hospital¹,
Pablo D. Dans¹ and Modesto Orozco^{1,2,*}

¹ Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain.

² Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain.

* To whom correspondence should be addressed: Prof. Modesto Orozco, Tel: +34 93 403 7155, Fax: +34 93 403 7157, Email: modesto.orozco@irbbarcelona.org.

SUPPORTING METHODS

The choice of sequences. We built a library of 40 different 16 bp oligomer sequences with a middle d(CpTpApG)₂ that cover the entire hexanucleotide space featuring a XpCpTpApGpX sequence pattern (X stands for any nucleotide) as well as all possible pyrimidine(Y)/purine(R) combinations at the octanucleotide level in several (>3) repeats.

System preparation and MD simulations. All the sequences were prepared with the leap program of AMBERTOOLS 16 (1) and simulated using pmemd.cuda code (2). Following the ABC protocol (3), canonical duplexes were generated using Arnott B-DNA fiber parameters (4), and solvated by a truncated octahedral box with a minimum distance of 10 Å between DNA and the closest face of the box.

Simulations were run using parmbsc1 force-field, SPC/E water model (5) and 150 mM concentration of K⁺Cl⁻ salt using Smith/Dang parameters (6–8). Systems were optimized and equilibrated as described in our previous works, and simulated for at least 500 ns and up to 10 μs in the NPT ensemble, using Particle-Mesh Ewald

corrections (2, 9) and periodic boundary conditions. SHAKE was used to constrain bonds involving hydrogen (10), allowing 2 fs integration step. All the trajectories and the associated analysis are accessible in the BigNASim portal: <https://mmb.irbbarcelona.org/BIGNASim/>.

Analysis of Molecular Dynamics trajectories. All the trajectories were processed with the *cptraj* module of the AMBERTOOLS 16 package (1), and the NAFlex server (10) for standard analysis. DNA helical parameters and backbone torsion angles were measured and analysed with the CURVES+ and CANAL programs (11), following the standard ABC conventions (3). The CANION module from Curves+ (12) was used to determine the position of cations in curvilinear cylindrical coordinates for each snapshot of the simulations with respect to the instantaneous helical axis. We obtained and analysed the ion distribution in one- (R, D, A) and two-dimensional (RA, DA, DR) curvilinear cylindrical coordinates at the central tetranucleotide sequence. Duplexes were named following the Watson strand (*e.g.* CTAG stands for (CTAG)-(CTAG)). The letters R, Y and X stand for a purine a pyrimidine or any base respectively, while X·X and XX represent a base pair and base-pair step respectively. Base pairs flanking the CTAG were denoted using two dots to represent the central tetrad (*e.g.* R··Y).

The Essential Modes of generic TpA in helical space. We performed Principal Component Analysis (PCA) of the 18 intra- and inter- base-pair parameters that define all degrees of freedom of the central TpA step in a rigid-base model. Before calculating the covariance matrix in helical space, its entries had to be made dimensionally uniform, so all rotational degrees of freedom were scaled by a factor of 10.6 (13). The covariance was calculated from the joint equilibrated trajectories of all 40 sequences taken at every 100 ps. The first 3 Principal Components, which explain ~60% of the total variance, have their largest projections on a subset of 8 of the original 18 helical parameters. These 3 PCs were used to perform multidimensional clustering in the essential helical space using the mclust package of R. The clustering is performed using the optimal model according to Bayesian Information Criterion (BIC) for an expectation-minimization (EM) algorithm initialized by hierarchical clustering for parameterized Gaussian mixture models.

Distributions of helical parameters that guide specific sequence dependence. The helical parameters that showed the highest variability across trajectories of different sequences were identified using Principal Component Analysis (PCA) of the 18 intra- and inter base pair parameters that define all degrees of freedom of the central TpA step in a rigid-base model. The first 3 Principal Components, which explain ~60% of the total variance have their largest projections on a subset of 8 of the original 18 helical parameters. The Bayesian Information Criterion (BIC) (14, 15) was used, limiting the analysis to either two or three components to determine the number of normal functions needed to meaningfully represent the appearance

of possible substates in the shift, slide, roll and twist 1D distributions of the joint trajectory of all sequences. The normal distributions obtained from the BIC decomposition were compared to the distributions of the same parameters obtained after the multivariate clustering (into 3 clusters) of the first 3 PCs.

From the eight parameters identified from the PCA as accounting for the most variance, six are non-collinear in the essential helical space, namely the shift, slide and twist of TpA bps, the buckle and propeller twist of dT and the buckle of dA. The distributions of the subset of these 6 parameters were used to evaluate the similarity between central TpA steps in different oligonucleotide sequences using the Kullback-Leibler (KL) divergence theorem. For each pair of oligomers we calculated the symmetrized values of the KL divergence and then applied hierarchical cluster analysis using Ward's clustering criterion (16), where the dissimilarities are squared before cluster updating (17) in order to identify specific sequence effects on TpA helical space flexibility.

The 4-state model of TpA dynamics. The 3D and 2D distributions of these three parameters and their paired combinations, respectively, in the meta-trajectory have also been calculated and they show a clear preference of the TpA to occupy one of four states in the Shift-Slide-Twist space. In fact, the states of the 3 helical parameters that display polymorphisms are highly inter-dependent, as shown in the 2- and 3- dimensional distribution plots. The 3 most populated states in the twist-slide-shift space, when considering the entire meta-trajectory of all oligonucleotides, are: High Twist/Positive Slide/Negative Shift (HPN), High Twist/Positive Slide/Positive Shift (HPP), and Low Twist/Negative Slide/Zero Shift (LNZ). In order to capture and better understand these effects, we filtered the meta-trajectory into 3 sub-trajectories corresponding to the 3 states, removing all frames that did not belong to any of these. We compared the distribution of helical parameters beyond the next-to-nearest neighbours (octanucleotide level) in both directions ("-" sign for moving towards the 5' direction on the Watson strand and "+" sign for the 3' direction) between the 3 substate-trajectories and found significant effects in the neighbouring shift, slide and twist. We also compared up to the octanucleotide level, backbone torsions, sugar puckering, and glycosidic torsions.

Breaking down the twist, slide and shift contributions to the distal sequence effects, we calculate the Pearson's correlations of these parameters at TpA to the helical parameters at one and two levels away from TpA in each direction and the point biserial correlations to the backbone torsion (zeta - categorized in trans and gauche-), sugar pucker (categorized into South and North) and glycosidic torsion (categorized into Anti and High Anti).

Equilibrium distributions of inter base pair helical parameters at the TpA step vary beyond next-to-nearest neighbours. BIC (Bayesian Information Criterion) was used to distinguish between the normal (one Gaussian) or multi-normal (a mixture of two or more Gaussians) nature of the distributions of TpA helical parameters (14, 15).

Since for each individual trajectory, the BIC decomposition assign the same number of Gaussians (1, 2 and 3) in the respective helical parameters (roll, twist/slide and shift, respectively) and the peaks of the distributions are consistent thought the set of oligomers, we compare the propensities of each Gaussian of the individual trajectories with the total average propensity per peak, assigning them to one of three ranges: mean – sd, mean + sd and within this interval, in order to identify large deviations in population imposed by sequence.

Correlation between twist and zeta states. As previously analysed in depth for the CpG case, we found strong correlations between the twist state and the BI/BII backbone state at the 3' side of the TpA step on both Watson and Crick strands. The backbone state was defined by discretizing the zeta torsion sub-states into trans (180 ± 40 degrees – associated with a backbone in BII), gauche positive (60 ± 40 degrees – extremely infrequent) and gauche negative (300 ± 40 degrees – associated with a backbone in BI). Just like in the CpG case, a low twist state was found to usually be coupled with BII transitions at both 3' junctions.

Correlation between twist and C-H··O3' hydrogen bond. Relying on strong evidence from previous studies (18, 19) of almost perfect correlation between backbone state and the formation of base to backbone hydrogen bonds, we looked at the correlation between twist state at the TpA step and hydrogen bond formation beyond the next-to-nearest neighbours. We found, as expected, a dependency of 3' side adjacent bond formation to twist state that perfectly mirrors that of the backbone state. But we also discovered an insightful sequential anti-correlation of bond formation from one step to the next that is also highly dependent on sequence, which favours the formation of one or the other.

Stacking and Base-pairing strength. In order to estimate the strength of stacking at the TpA step we calculated a Stacking Factor based on the distance between the centres of mass of DT and DA, and the angle between the two planes of the bases, defined as (20):

$$\xi = \frac{r_M}{S(\alpha)}$$

$$S(\alpha) = e^{-\alpha} + e^{-(\alpha-\pi)^4} + 0.1e^{-(\alpha-0.5\pi)^4}$$

where r_M is the distance between the two centres of mass and α the angle between the base planes. We calculated the Stacking Factors separately for the major 3 of the 4 states in twist/slide/shift space defined above to determine the stabilizing factors of the highly preferred states.

Database Analysis of structural features. We retrieved high resolution ($< 3\text{\AA}$) structures of double stranded DNA containing the CTAG tetrad and distinguished between the protein-bound and free DNA structures. We compared helical parameter distributions and components of BIC analysis between the database structures and our results. We paid special attention to the sequence context bias found in the database and performed the comparison to the meta-trajectory from simulations containing the same hexanucleotide environments centred at TpA.

Database Analysis of genomic properties. Prevalence of CTAG in the genomes of *H. sapiens* (hg19), *E. coli* (NC_000913.3) and *S. cerevisiae* (sacCer3) was computed, finding low occurrence compared to other tetranucleotides (less than 0.5% in the three species). Occurrences of this tetranucleotide were then mapped, using Homer software (21), to the annotated regions of each organism obtained from UCSC and compared to the overall frequency of each annotation type. CTAG is enriched at intergenic regions in *H. sapiens* and *E. coli*, but not in *S. cerevisiae* probably due to the low number of intergenic regions in this organism (less than 2.5% compared to more than 20% in the other two). To evaluate resilience to mutation, the frequency of mutations for each tetranucleotide (normalised by tetranucleotide frequency) along the genome in 30 different cancer types (22) was computed. SNPs in human genome were retrieved from Ensembl Variation database (23) and were mapped to each tetranucleotide to compute normalized SNP frequency per tetranucleotide.

SUPPORTING TABLES

Table S1. Sequence library used to study CTAG polymorphisms, number of replicas and simulation time.

Num.	Sequence	Simulation time	Num.	Sequence	Simulation time
1	CGTCGGCTAGCCGAGC	500 ns	21	CGGAGACTAGACTCGC	500 ns
2	CGTCTCCTAGGAGAGC	500 ns	22	CGGAGACTAGCCTCGC	500 ns
3	CGAAAACTAGAAAAGC	500 ns	23	CGGAGACTAGGCTCGC	6 μ s
4	CGAAAACTAGTTTTGC	500 ns	24	CGGAGACTAGTCTCGC	6 μ s
5	CGATATCTAGATATGC	500 ns	25	CGGAGCCTAGACTCGC	500 ns
6	CGTATACTAGTATAGC	2 x 500 ns	26	CGGAGCCTAGCCTCGC	2 x 500 ns
7	CGGGGGCTAGGGGGGC	500 ns	27	CGGAGCCTAGGCTCGC	500 ns
8	CGGGGGCTAGCCCCGC	500 ns	28	CGGAGGCTAGACTCGC	500 ns
9	CGGCGCCTAGGCGCGC	500 ns	29	CGGAGGCTAGCCTCGC	6 μ s
10	CGCGCGCTAGCGCGGC	500 ns	30	CGGAGTCTAGACTCGC	2 x 500 ns
11	CGTCTACTAGAGAGGC	500 ns	31	CGCTAGCTAGCTAGGC	4 x 500 ns
12	CGTCTACTAGCGAGGC	2 x 500 ns	32	CGATATCTAGAAATGC	2 μ s
13	CGTCTACTAGGGAGGC	6 μ s	33	CGGAGCCTAGAATCGC	2 μ s
14	CGTCTACTAGTGAGGC	2 x 500 ns	34	CGGCGCCTAGGGGGCGC	2 μ s
15	CGTCTCCTAGAGAGGC	2 x 500 ns	35	CGGAGGCTAGCATCGC	2 μ s
16	CGTCTCCTAGCGAGGC	500 ns	36	CGAAAACTAGTATAGC	2 μ s
17	CGTCTCCTAGGGAGGC	500 ns	37	CGCTAGCTAGCGAGGC	2 μ s
18	CGTCTGCTAGAGAGGC	6 μ s	38	CGTCTGCTAGACAGGC	2 μ s
19	CGTCTGCTAGCGAGGC	9 μ s	39	CGAATCCTAGATAAGC	2 μ s
20	CGTCTTCTAGAGAGGC	500 ns	40	CGGACACTAGCGTCGC	2 μ s

Table S2. Pearson correlation coefficients of Shift, Slide and Twist at TpA with flanking bps parameters and selected backbone torsions up to next-to-nearest neighbours.

		Shift at TA	Slide at TA	Twist at TA		Shift at TA	Slide at TA	Twist at TA
-2	Shift	0.06	0.002	0.025	zetaW	-0.067	-0.063	-0.123
	Slide	0.157	0.149	0.206	zetaC	-0.471	-0.286	-0.421
	Rise	-0.052	-0.022	-0.086	phaseW	-0.130	-0.023	-0.073
	Tilt	0.086	0.031	0.051	phaseC	-0.061	-0.079	-0.110
	Roll	0.001	0.043	0.038	chiW	0.018	0.002	0.025
	Twist	0.089	0.051	0.021	chiC	-0.074	-0.042	-0.057
-1	Shift	-0.607	-0.149	-0.257	zetaW	-0.454	-0.098	-0.217
	Slide	-0.298	0.089	-0.094	zetaC	0.753	0.295	0.536
	Rise	0.028	-0.089	-0.109	phaseW	-0.425	0.006	-0.105
	Tilt	-0.12	0.057	-0.11	phaseC	0.111	0.102	0.090
	Roll	0.002	0.178	0.157	chiW	-0.140	-0.027	-0.058
	Twist	-0.223	-0.263	-0.453	chiC	0.107	0.173	0.153
Central TpA step								
+1	Shift	-0.607	0.192	0.306	zetaW	-0.736	0.340	0.589
	Slide	0.201	0.098	-0.078	zetaC	0.456	-0.166	-0.260
	Rise	0.017	-0.08	-0.114	phaseW	-0.157	0.130	0.103
	Tilt	-0.104	-0.047	0.12	phaseC	0.431	-0.045	-0.144
	Roll	-0.045	0.176	0.173	chiW	-0.206	0.186	0.170
	Twist	0.232	-0.25	-0.455	chiC	0.166	-0.022	-0.053
+2	Shift	0.185	-0.084	-0.148	zetaW	0.547	-0.332	-0.487
	Slide	-0.251	0.195	0.271	zetaC	0.023	-0.023	-0.061
	Rise	0.09	-0.04	-0.103	phaseW	0.020	-0.072	-0.076
	Tilt	0.156	-0.091	-0.125	PhaseC	0.085	-0.004	-0.054
	Roll	0.012	0.044	0.039	chiW	0.019	-0.012	-0.018
	twist	-0.095	0.079	0.067	chiC	-0.067	0.006	0.024

Table S3. Number and frequency of unique occurrences of hexanucleotides containing central CTAG in the PDB database.

Type	Hexanucleotide Context	No. structures	Frequency
Naked DNA structures	G..C	15	0.54
	A..T	5	0.18
	T..A	3	0.11
	C..G	2	0.07
	T..T	2	0.07
	T..C	1	0.04
	A..G	30	0.31
	G..A	30	0.31
Protein-DNA complexes	T..A	11	0.11
	A..T	8	0.08
	G..G	7	0.07
	C..G	5	0.05
	A..A	2	0.02
	G..C	2	0.02
	A..C	1	0.01
	T..G	1	0.01

SUPPORTING FIGS.

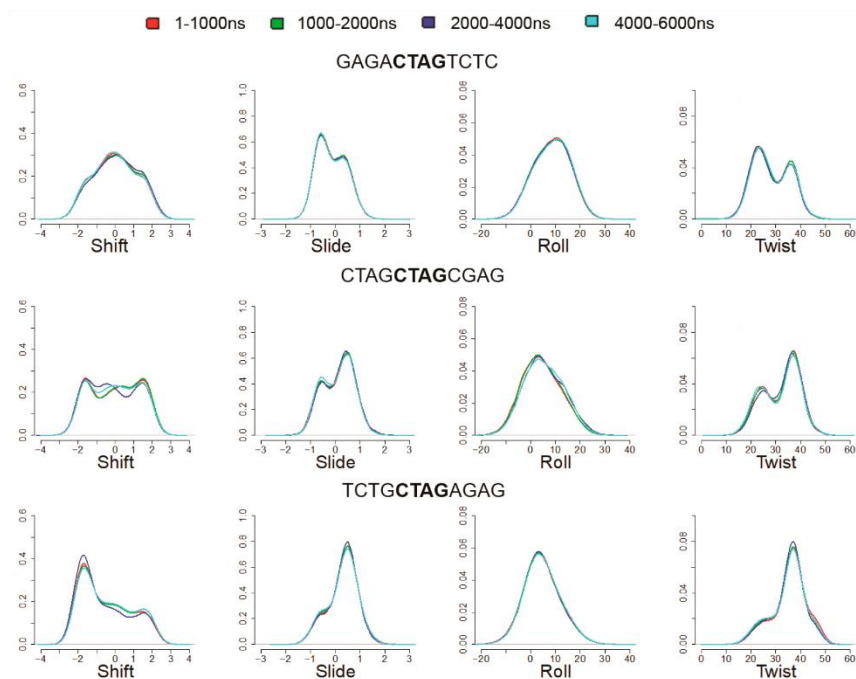


Fig. S1. Normalized frequencies of the shift, slide, roll and twist helical parameters for 3 selected sequences, whose trajectories were extended to 6 μ s to check for convergence. Four distributions were computed for each helical parameter using segments of 1,000 or 2,000 ns.

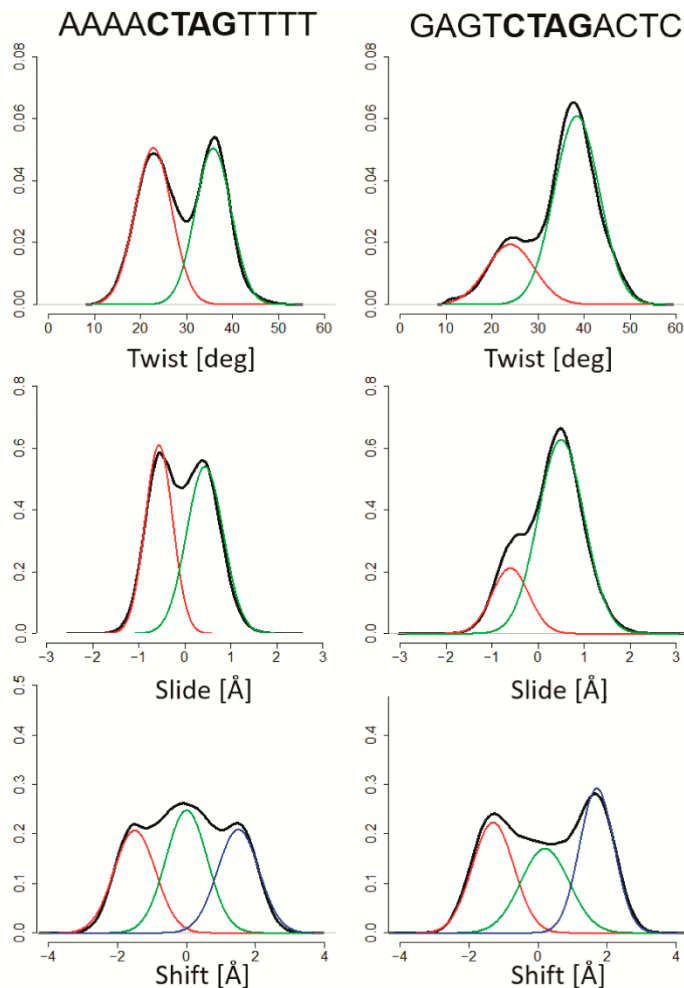


Fig. S2. Normalised frequencies of the shift, slide, and twist helical parameters for 2 selected sequences showing clear next-to-nearest neighbour effects, which could be appreciated from the change in the relative populations of the bi- and tri-normal distributions.

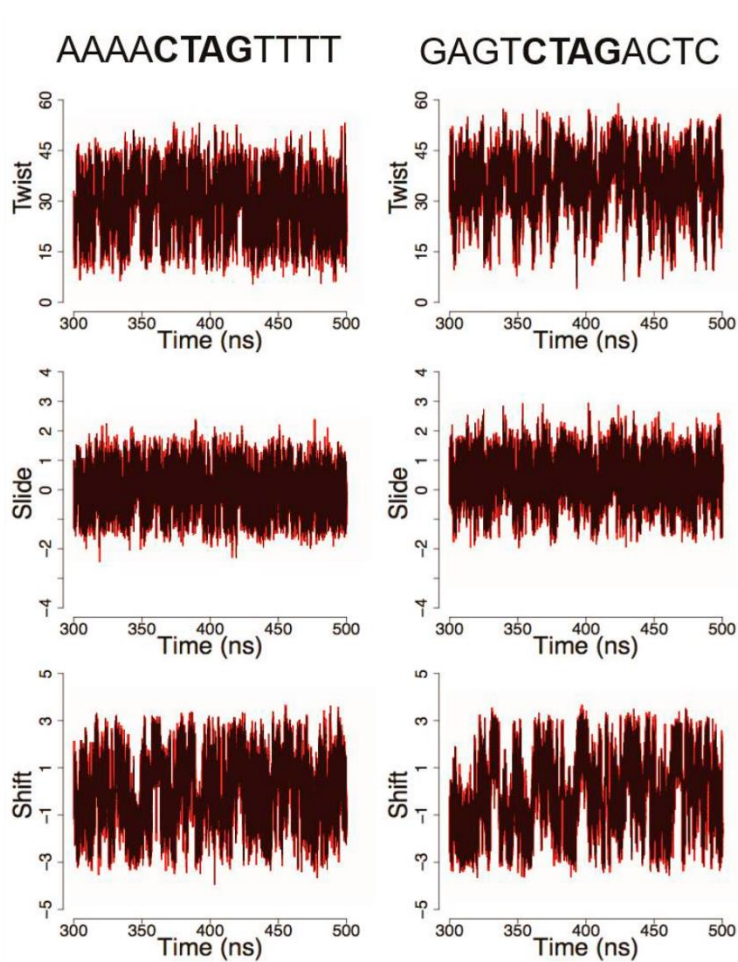


Fig. S3. Time evolution (500 ns) of shift, slide and twist for two selected sequences, showing the fast and reversible inter-conversion between high and low substates.

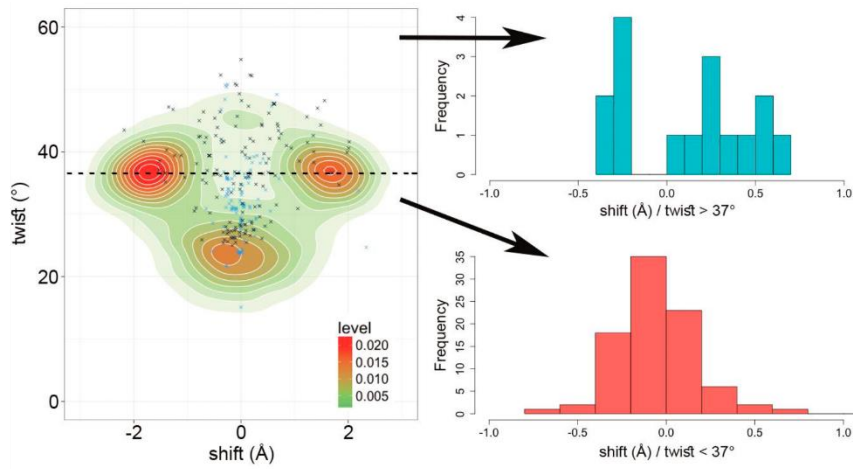


Fig. S4. 2D counts in the shift-twist plane from MD simulations at the central TpA step. In the 2D density plots experimental structures from the PDB (see Supp. Methods) were added as black crosses (Protein-DNA complexes), or blue crosses (isolated DNA). We divided the plane between high twist ($> 37^\circ$), and low twist ($< 37^\circ$) and analysed the shift distribution for these two cases.

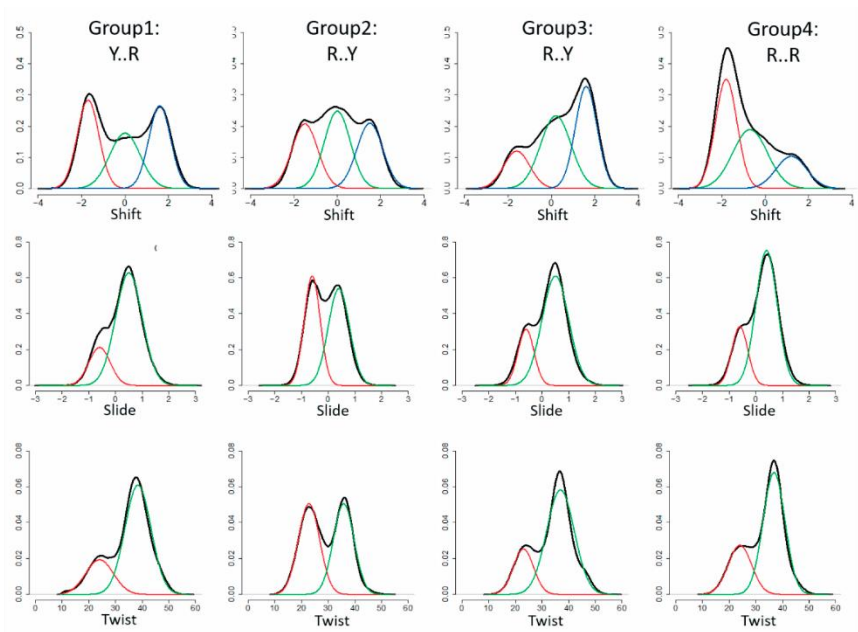


Fig. S5. Normalized frequencies for shift, slide and twist (black line), and the BIC decomposition in Gaussians (red, green, and blue lines), showing the behaviour of the clusters obtained in the dendrogram of Fig. 5.

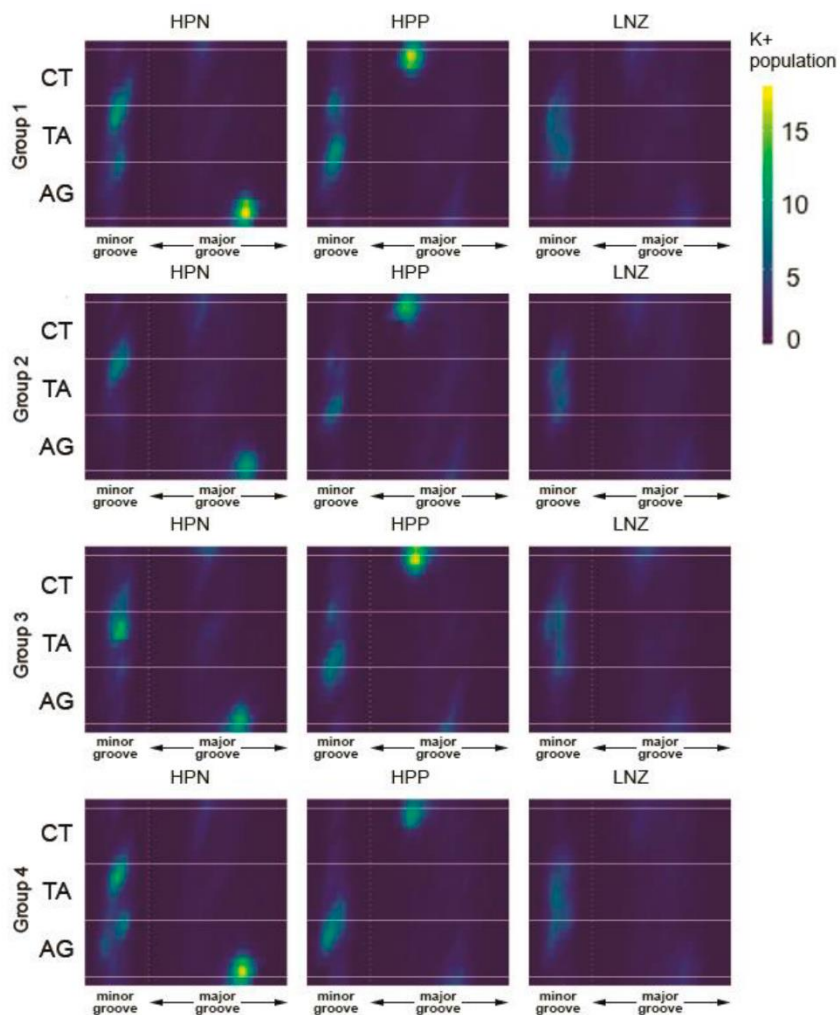


Fig. S6. Population of K⁺ ions inside the major and minor groove for bps CpT, TpA, and ApG in each of the three major states based on twist/slide/shift values at TpA.

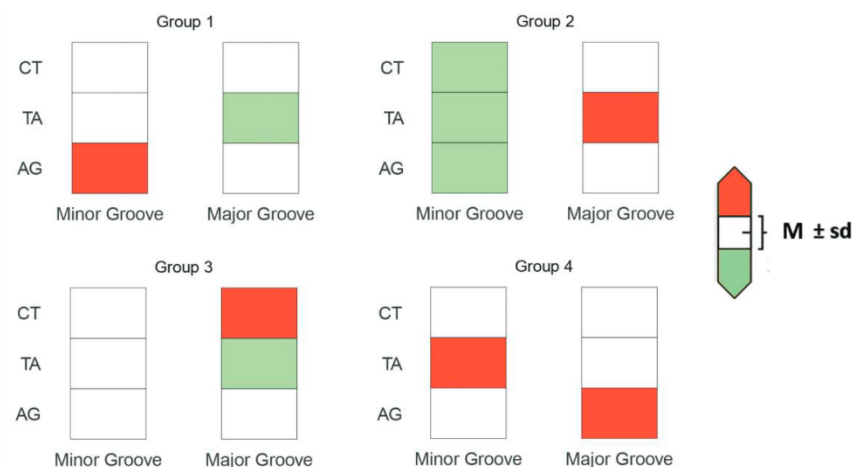


Fig. S7. Relative ion populations of cluster representatives in the minor and major groove at the CTAG tetranucleotide. Comparison to the global average ion populations per region.

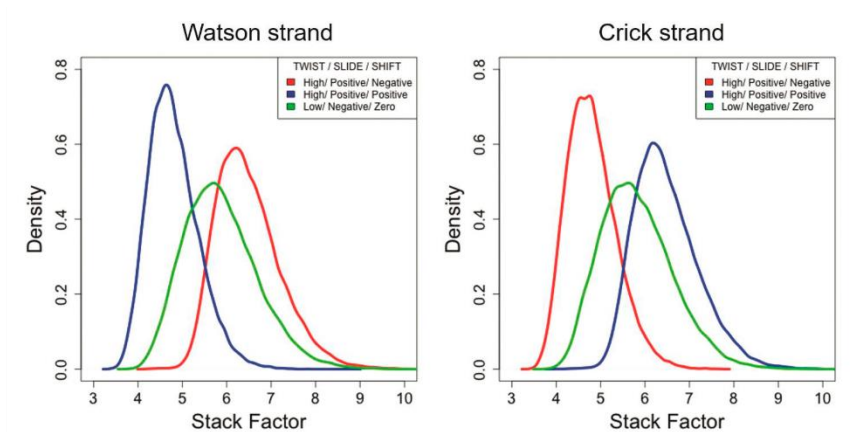


Fig. S8. Distributions of stacking coordinate at the TpA step on both Watson (left) and Crick (right) strands in the three main configurations of the bps in helical space.

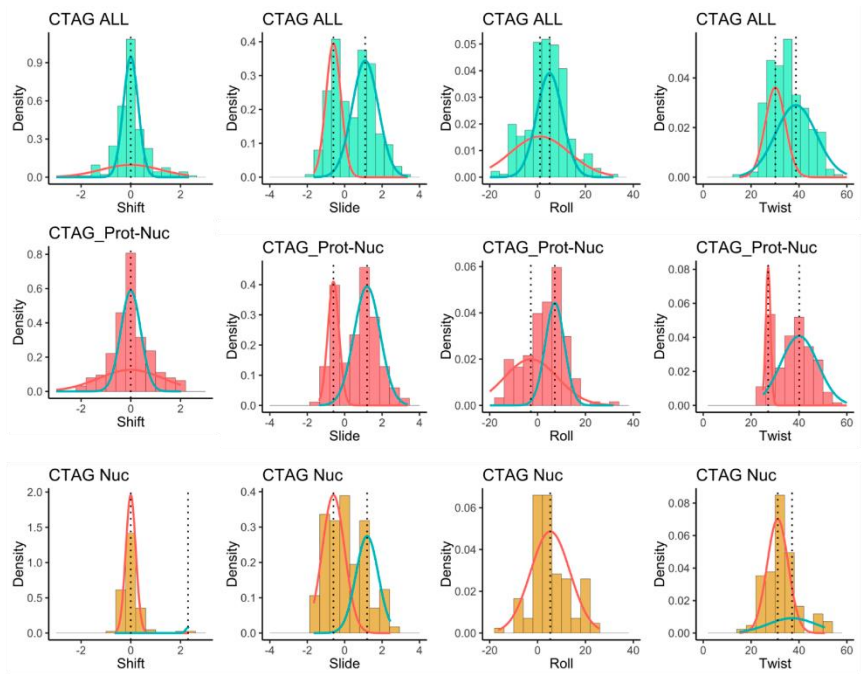


Fig. S9. Normalized frequencies of shift, slide, roll and twist at TpA obtained from the data mining of the PDB for all structures containing CTAG according to BIC analysis: all DNA (FIRST ROW), Protein-DNA complexes (SECOND ROW), and isolated DNA structures (THIRD ROW).

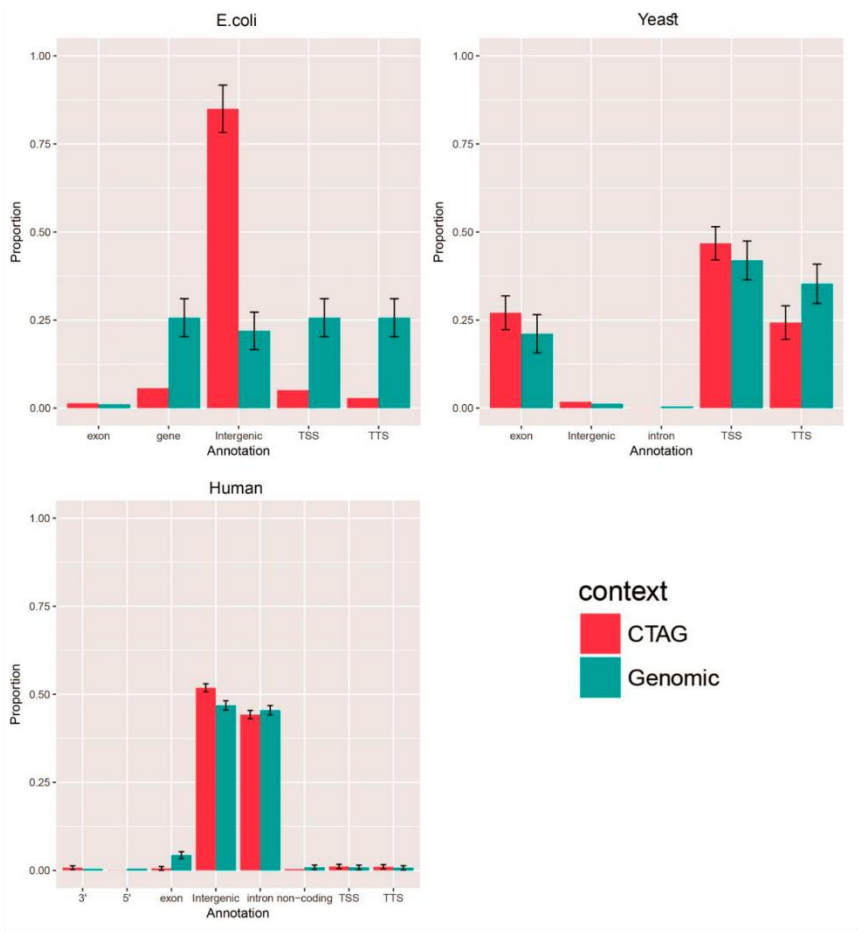


Fig. S10. Occurrence of CTAG in different genomic regions. Length of each annotation type is shown to evaluate significance of enrichment per region type.

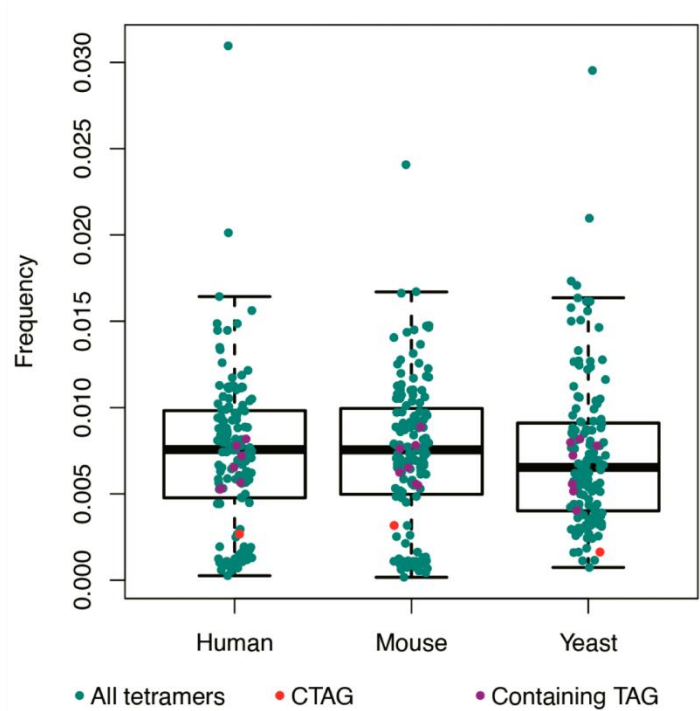


Fig. S11. Frequency of each possible tetranucleotide in 3 different genomes. CTAG is marked in red, tetranucleotides containing TpApG (all but the amber stop codon) are marked in violet, and the rest are depicted in cyan. Note that this analysis doesn't includes exons.

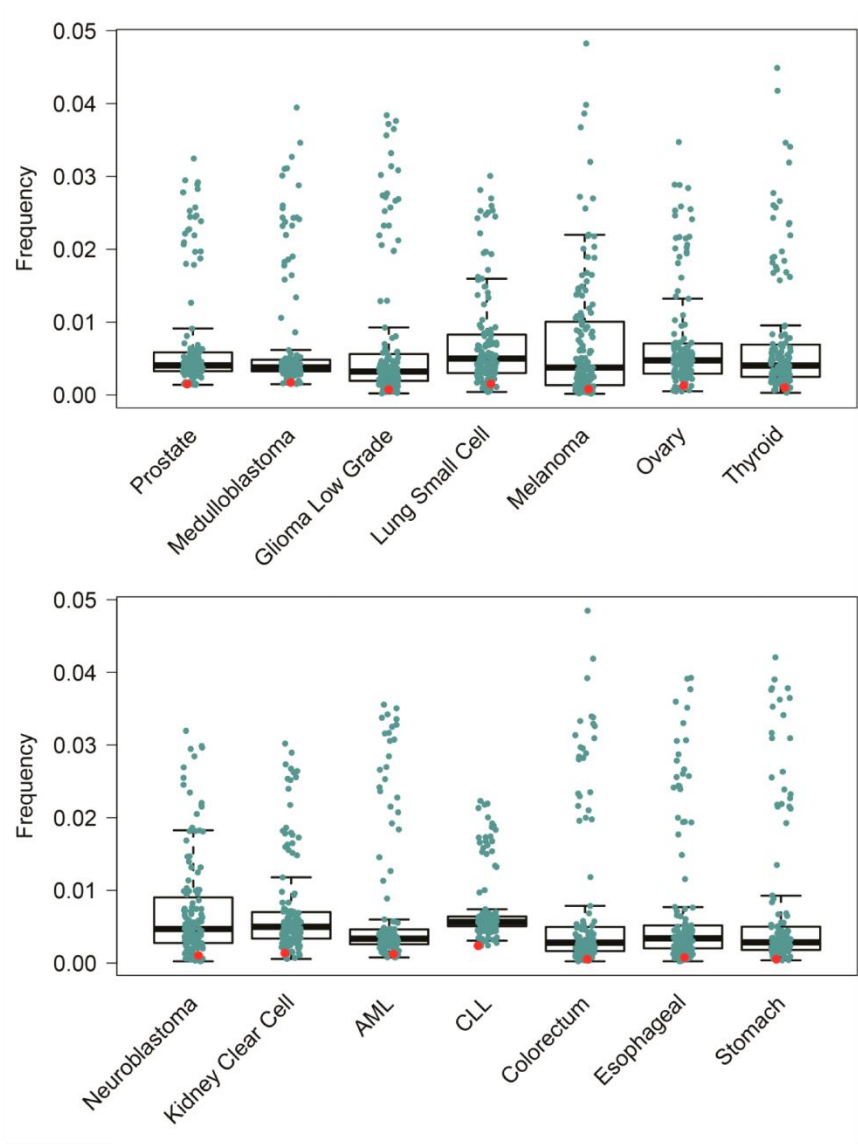


Fig. S12. Frequency of mutations for each tetranucleotide along the genome for several cancer types, normalised by genome-wide tetranucleotide occurrence. CTAG is marked in red.

SUPPORTING REFERENCES

1. D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, L.X. and P.A.K. (2016) AMBER 2016.
2. Le Grand, S., Götz, A.W. and Walker, R.C. (2013) SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.*, **184**, 374–380.
3. Pasi, M., Maddocks, J.H., Beveridge, D., Bishop, T.C., Case, D.A., Cheatham, T., Dans, P.D., Jayaram, B., Lankas, F., Laughton, C., *et al.* (2014) μ ABC: A systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
4. Arnott, S. and Hukins, D.W.L. (1973) Refinement of the structure of B-DNA and implications for the analysis of X-ray diffraction data from fibers of biopolymers. *J. Mol. Biol.*, **81**, 93–105.
5. Berendsen, H.J.C., Grigera, J.R., Straatsma, T.P., Grigera, J.R., Straatsma, T.P., Berendsen, H., Grigera, J., Straatsma, T., Grijera, J., Berendsen, H.J.C., *et al.* (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
6. Smith, D.E. and Dang, L.X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.*, **100**, 3757–3766.
7. Dang, L.X. (1995) Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
8. Dang, L.X. and Kollman, P.A. (1995) Free Energy of Association of the K^+ :18-Crown-6 Complex in Water: A New Molecular Dynamics Study. *J. Phys. Chem.*, **99**, 55–58.
9. Darden, T., York, D. and Pedersen, L. (1993) Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.
10. Ryckaert, J.-P., Ciccotti, G. and Berendsen, H.J. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.
11. Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
12. Pasi, M., Maddocks, J.H. and Lavery, R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2412–23.
13. Dršata, T. and Lankaš, F. (2013) Theoretical models of DNA flexibility. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 355–363.
14. Schwarz, G. (1978) Estimating the Dimension of a Model. *Ann. Stat.*, **6**, 461–464.
15. Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
16. Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *J. Am.*

- Stat. Assoc.*, **58**, 236–244.
17. Murtagh, F. and Legendre, P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.*, **31**, 274–295.
 18. Dans, P.D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R. and Orozco, M. (2014) Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res.*, **42**, 11304–11320.
 19. Balaceanu, A., Pasi, M., Dans, P.D., Hospital, A., Lavery, R. and Orozco, M. (2017) The Role of Unconventional Hydrogen Bonds in Determining BII Propensities in B-DNA. *J. Phys. Chem. Lett.*, **8**.
 20. Jafilan, S., Klein, L., Hyun, C. and Florián, J. (2012) Intramolecular Base Stacking of Dinucleoside Monophosphate Anions in Aqueous Solution. *J. Phys. Chem. B*, **116**, 3613–3618.
 21. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.
 22. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A. V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
 23. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

II. Sequence selective protein-DNA recognition

SUPPLEMENTARY MATERIAL

How B-DNA dynamics decipher sequence-selective protein recognition

Federica Battistini¹, Adam Hospital¹, Diana Buitrago¹,

Diego Gallego¹, Pablo D. Dans¹, Josep Lluís Gelpi² and Modesto Orozco^{1,2,*}

¹ Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain.

² Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain.

Supplementary Method

PDB Protein-DNA dataset filtering. The dataset representing the DNA-protein interactome used in the study was obtained after applying a set of filters to the whole repository of protein complexes deposited in the Protein Data Bank. The initial dataset was acquired from Nucleic Acid Database (NDB)[1] [accessed on Feb-2018], selecting PDB entries having protein molecules attached to double-stranded B-DNA, thus avoiding single-stranded nucleic acid structures, RNA, and non-canonical B-DNA conformations. From this dataset of 3,465 different entries, single protein-DNA complexes were extracted in the cases where more than one existed in the PDB file (*e.g.* 1PP8). From the resulting list, complexes with DNA molecules having one of the following issues were removed:

- 1) Modified nucleotides or non-standard bases. Information taken from PDBe REST API [<http://www.ebi.ac.uk/pdbe/pdbe-rest-api>]. (defined as Modified/Broken)
- 2) Duplex strands broken, base pair mismatch or a malformed Watson-Crick base pairing (evaluated from the hydrogen bonds formed between the two bases) (defined as Mismatch in this work).
- 3) Unpaired strands having over-hanging bases with no associated pair (defined as Unpaired in this work).
- 4) Highly distorted-unnatural double-stranded B-DNA (having negative values for Twist and/or Rise helical parameters) (Unnatural_DNA).

The final dataset is composed of 174 no redundant protein-DNA complexes.

Simulation conditions. All simulations were performed in the isothermal-isobaric ensemble (NPT, P=1 atm and T=298 K) applying the Berendsen algorithm[2] to control the temperature and the pressure, with a coupling constant of 5 ps and removing center of mass motion every 10 ps. Long-range electrostatic interactions were accounted for by using the Particle Mesh Ewald method[3] with standard defaults and a real-space cutoff of 10 Å. All bonds involving hydrogen were kept constrained at equilibrium values using SHAKE[4], which allowed us to use a 2 fs step for the integration of Newton’s equations of motion. The DNA interactions were represented using the new Parmbsc1 force-field[5–7] All simulations were performed using Amber 14 suite of programs (AMBER 2014 San Francisco University of California).

Relative Position: Relative position of the Deformation energy calculated for the DNA sequence extracted from the Xray crystal structure (def.Energy xray) respect to energy distribution built using the randomly generate sequence.

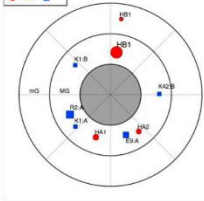
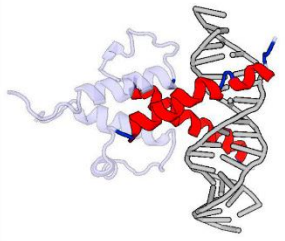
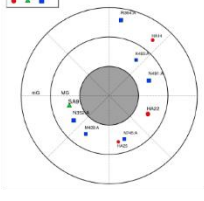

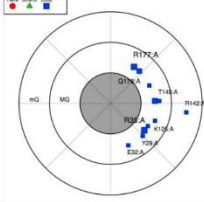
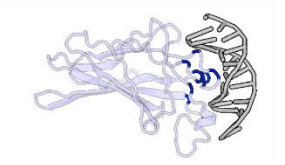
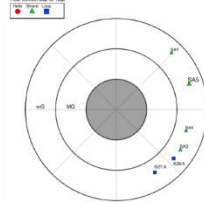

$$\text{Relative position (\%)} = \frac{\text{Def. Energy(xray)} - \text{Def. Energy (min)}}{\text{Def. Energy(max)} - \text{Def. Energy(min)}}$$

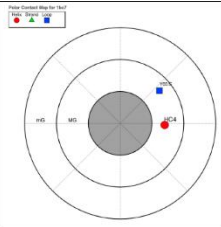
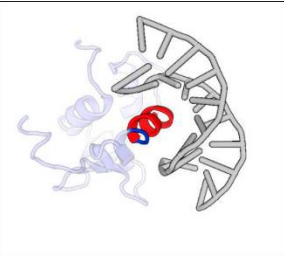
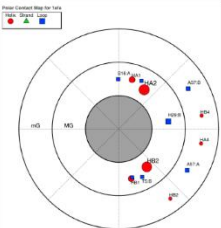
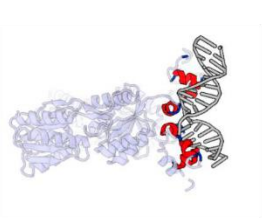
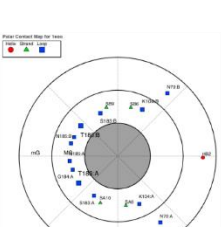
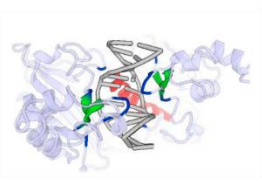
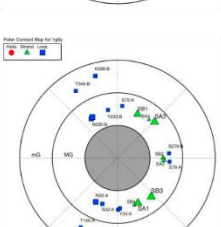
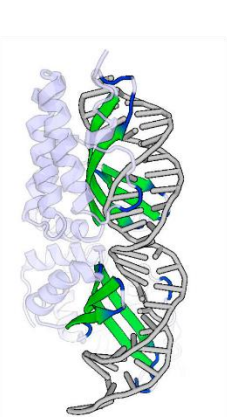
Where Def.Energy(max) and Def.Energy(min) are the maximum and minimum energy values respectively extracted from the energy distribution built using randomly generated DNA sequences of the same length as the sequence in the Xray crystal structure.

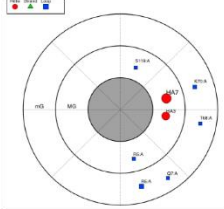
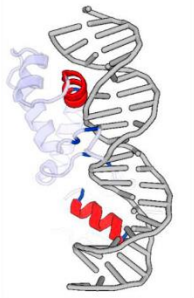
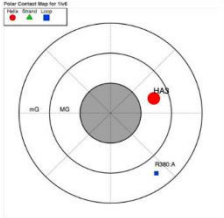
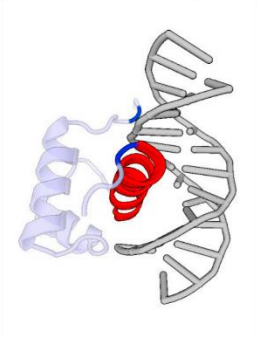
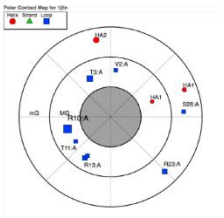
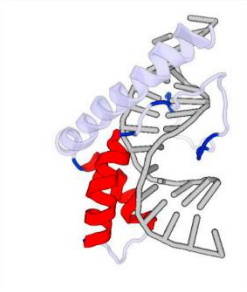
Supplementary Data

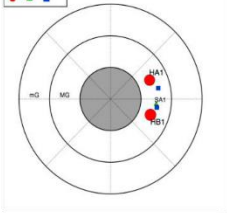
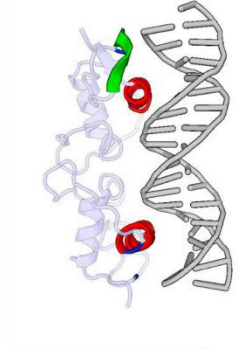
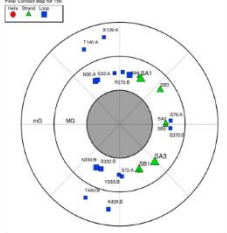

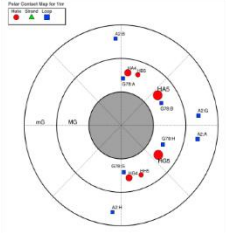
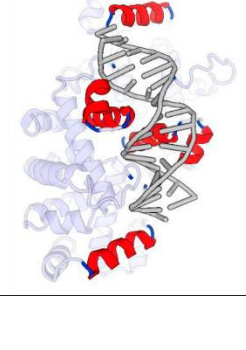
Supplementary Table S1 Summary of the PDB ID code and characteristics of the DNA-protein complexes selected.

PDB ID code	Biological function of the protein, classification Uniprot-GO and Experimental method ¹	Protein-DNA Binding Recognitin site ²	Representation ³
-------------	--	--	-----------------------------

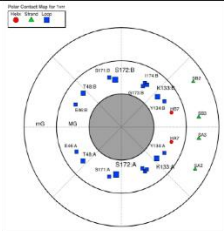
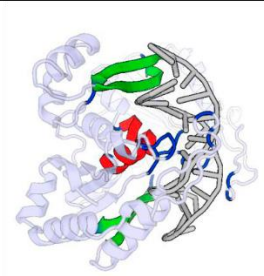
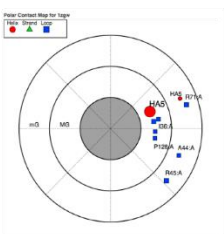
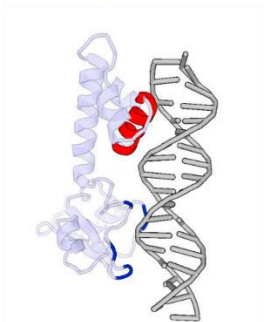
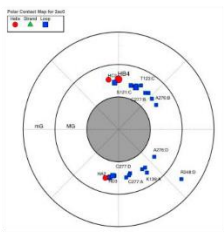
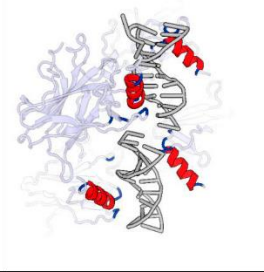
1A0A	Phosphate positive system protein PHO4, TF activity	<div><div>Protein Contact Map for 1A0A</div><div><div>1A0A</div><div>1A0A</div><div>1A0A</div></div></div>  <div>Xray 2.8Å</div>	
1A36	DNA topoisomerase 1	<div><div>Protein Contact Map for 1A36</div><div><div>1A36</div><div>1A36</div><div>1A36</div></div></div>  <div>Xray 2.8Å</div>	
1A66	Nuclear factor of activated T-cells, cytoplasmic 1, TF activity	<div><div>Protein Contact Map for 1A66</div><div><div>1A66</div><div>1A66</div><div>1A66</div></div></div>  <div>NMR (18)</div>	
1AZP	DNA-binding protein 7d	<div><div>Protein Contact Map for 1AZP</div><div><div>1AZP</div><div>1AZP</div><div>1AZP</div></div></div>  <div>Xray 1.6Å</div>	
1BC7	ETS domain-containing protein Elk-4, TF activity		

	Xray 2.0Å		
1EFA	Lactose operon repressor TF repressor activity Xray 2.6Å		
1E00	Type-2 restriction enzyme EcoRV Xray 2.16Å		
1G9Y	DNA endonuclease I-CreI Xray 2.05Å		

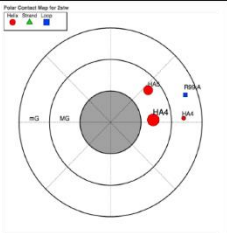
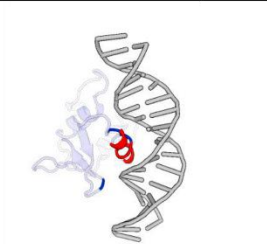
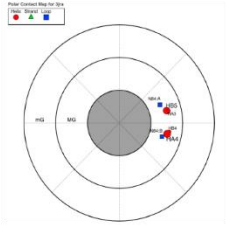
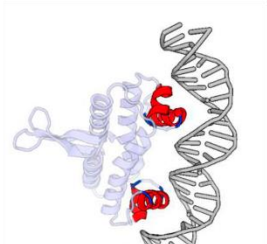
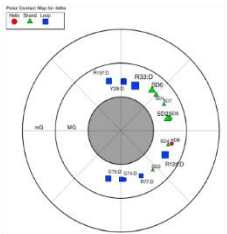
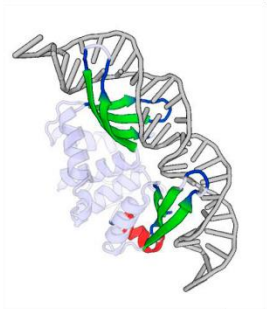
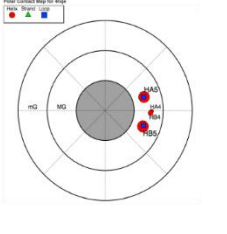
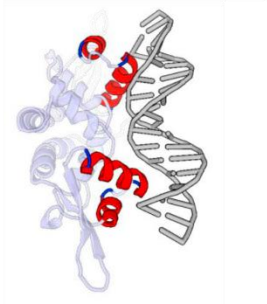
1HLV	Major centromere autoantigen B Xray 2.5Å	 <p>Polar Contact Map for 1HLV 100% 50% 0% Red Green Blue</p> <p>This polar contact map shows interactions between residues 1-100. The map is circular with concentric rings representing different residue ranges. The legend indicates that red represents 100% contact, green represents 50%, and blue represents 0%. The map shows several clusters of high contact, particularly in the 1-100 range.</p>	 <p>3D structure of Major centromere autoantigen B, showing the protein structure in red and blue, bound to DNA in grey.</p>
1IV6	Telomeric repeat-binding factor 1, TF activity NMR (20)	 <p>Polar Contact Map for 1IV6 100% 50% 0% Red Green Blue</p> <p>This polar contact map shows interactions between residues 1-100. The map is circular with concentric rings representing different residue ranges. The legend indicates that red represents 100% contact, green represents 50%, and blue represents 0%. The map shows a cluster of high contact in the 1-100 range.</p>	 <p>3D structure of Telomeric repeat-binding factor 1, showing the protein structure in red and blue, bound to DNA in grey.</p>
1J5N	Non-histone chromosomal protein 6A, TF activity NMR (20)	 <p>Polar Contact Map for 1J5N 100% 50% 0% Red Green Blue</p> <p>This polar contact map shows interactions between residues 1-100. The map is circular with concentric rings representing different residue ranges. The legend indicates that red represents 100% contact, green represents 50%, and blue represents 0%. The map shows several clusters of high contact, particularly in the 1-100 range.</p>	 <p>3D structure of Non-histone chromosomal protein 6A, showing the protein structure in red and blue, bound to DNA in grey.</p>

1R4I	Androgen receptor, TF activity Xray 3.1Å		
1T9I	DNA endonuclease I-Crel Xray 1.6Å		
1TRR	Trp operon repressor, TF activity Xray 2.4Å		

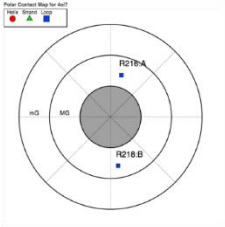
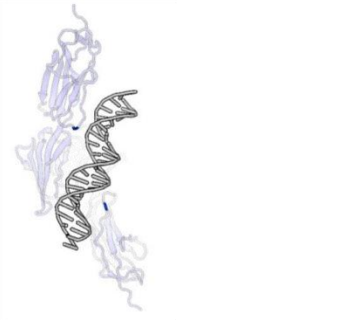
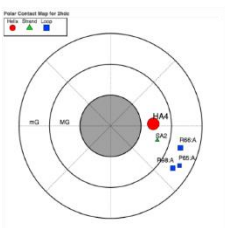
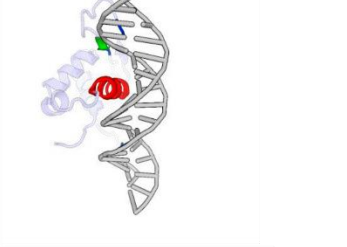
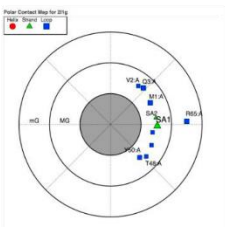
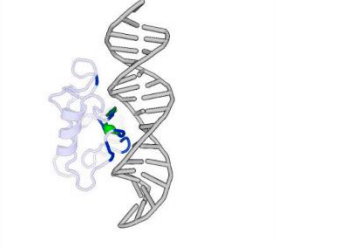
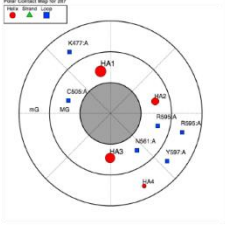
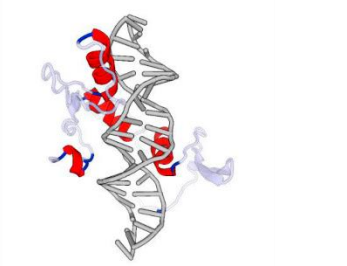
1UBD	Transcriptional repressor protein YY1, TF activity Xray 2.5Å	
------	---	--

1VRR	BstYI Xray 2.7Å		
1ZGW	Bifunctional transcriptional activator/DNA repair NMR		
2AC0	Cellular tumor antigen p53, TF activity Xray 1.8Å		

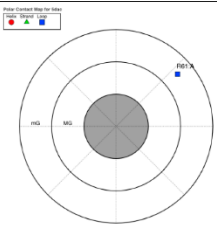
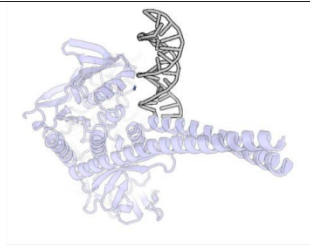
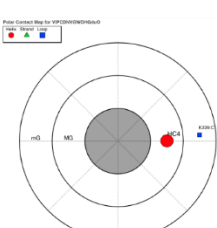

2GZK	Sex-determining region Y protein, TF activity NMR (1)		
2KDZ	MYB24 NMR (10)		
2QHB	Telomere binding protein TBP1 Xray 2.4Å		

2STW	Protein C-ets-1 TF activity NMR (1)	 <p>Protein Contact Map for 2STW Legend: HSA (red), HSB (green), HSC (blue)</p> <p>This circular contact map shows interactions between residues. The legend indicates HSA (red), HSB (green), and HSC (blue). The map shows several red and blue dots, indicating contacts between HSA and HSC residues. The axes are labeled with residue numbers: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000.</p>	 <p>3D ribbon diagram of Protein C-ets-1 (red) bound to a DNA double helix (grey). The protein is shown in a blue ribbon representation, highlighting its interaction with the DNA major groove.</p>
3JRA	DNA-binding protein Fis, TF activity Xray 3.11Å	 <p>Protein Contact Map for 3JRA Legend: HSA (red), HSB (green), HSC (blue)</p> <p>This circular contact map shows interactions between residues. The legend indicates HSA (red), HSB (green), and HSC (blue). The map shows several red and blue dots, indicating contacts between HSA and HSC residues. The axes are labeled with residue numbers: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000.</p>	 <p>3D ribbon diagram of Protein Fis (red) bound to a DNA double helix (grey). The protein is shown in a blue ribbon representation, highlighting its interaction with the DNA major groove.</p>
4D6O	Homing endonuclease I- Dmol Xray 2.2Å	 <p>Protein Contact Map for 4D6O Legend: HSA (red), HSB (green), HSC (blue)</p> <p>This circular contact map shows interactions between residues. The legend indicates HSA (red), HSB (green), and HSC (blue). The map shows several red and blue dots, indicating contacts between HSA and HSC residues. The axes are labeled with residue numbers: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000.</p>	 <p>3D ribbon diagram of Homing endonuclease I (red) bound to a DNA double helix (grey). The protein is shown in a blue ribbon representation, highlighting its interaction with the DNA major groove.</p>
4HQE	Quinone-sensing and response repressor QsrR, Uncharacterized protein Xray 2.29Å	 <p>Protein Contact Map for 4HQE Legend: HSA (red), HSB (green), HSC (blue)</p> <p>This circular contact map shows interactions between residues. The legend indicates HSA (red), HSB (green), and HSC (blue). The map shows several red and blue dots, indicating contacts between HSA and HSC residues. The axes are labeled with residue numbers: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000.</p>	 <p>3D ribbon diagram of Quinone-sensing and response repressor QsrR (red) bound to a DNA double helix (grey). The protein is shown in a blue ribbon representation, highlighting its interaction with the DNA major groove.</p>

3F27	Transcription factor SOX-17 Xray 2.75Å		
1H9T	Fatty acid metabolism, TF activity Xray 2.25Å		
1K60	ETS domain-containing protein Elk-4, TF activity Xray 3.19 Å		
1VTN	Hepatocyte nuclear factor 3-gamma, TF activity Xray 2.5Å		

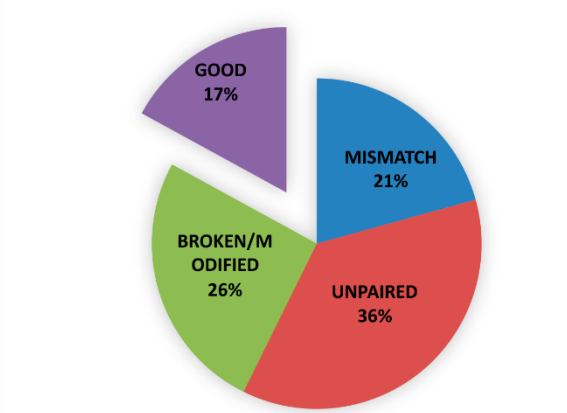
4017	Advanced glycosylation end product-specific Xray 3.1 Å		
2HDC	Forkhead box protein D3, TF activity NMR (20)		
2L1G	THAP domain-containing protein 1, TF activity NMR (17)		
2LT7	Transcriptional regulator Kaiso NMR (20)		

1CDW	TATA-box-binding protein TF activity Xray 1.9Å		
1J46	Sex-determining region Y protein, TF activity NMR (1)		
3U2B	Transcription factor SOX- 4 Xray 2.4Å		
5B7J	Switch-activating protein 1 NMR (20)		

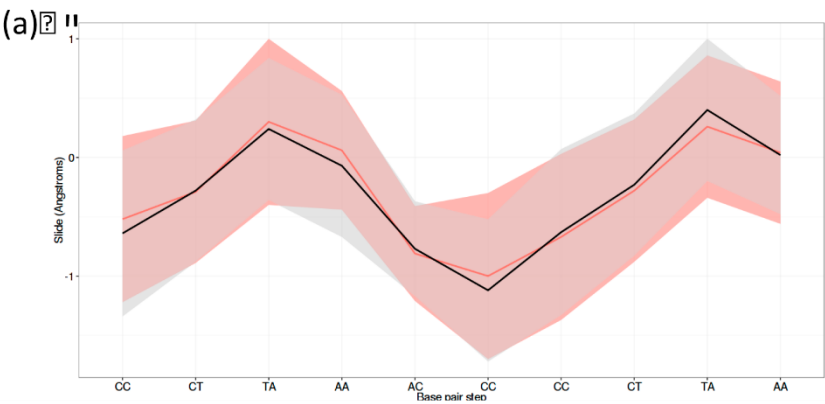
5DAC	ATPase activity Uncharacterized protein Xray 2.5Å	 <p>Polar Contact Map for 5DAC</p> <p>Legend: Major Groove (red), Minor Groove (green), Loop (blue)</p> <p>The map shows a single contact point in the major groove (red) at approximately 10 o'clock, labeled 'M1'.</p>	 <p>Cartoon representation of the 5DAC protein-DNA complex. The protein is shown in light blue, and the DNA is shown in grey.</p>
5X07	Hepatocyte nuclear factor 3-beta, TF activity Xray 2.79Å	 <p>Polar Contact Map for 5X07</p> <p>Legend: Major Groove (red), Minor Groove (green), Loop (blue)</p> <p>The map shows two contact points: one in the major groove (red) at approximately 10 o'clock, labeled 'M1', and one in the minor groove (green) at approximately 2 o'clock, labeled 'M2'.</p>	 <p>Cartoon representation of the 5X07 protein-DNA complex. The protein is shown in light blue, and the DNA is shown in grey.</p>

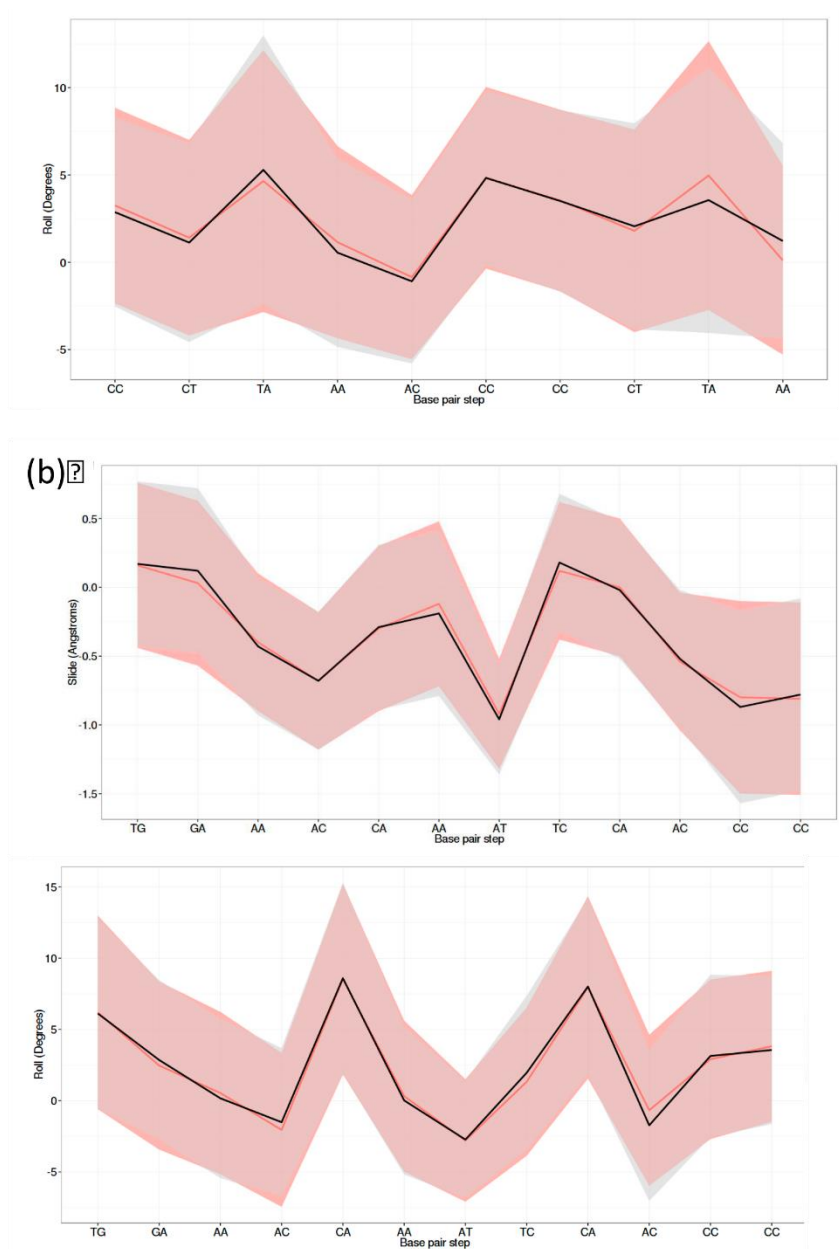
¹The biological function of the protein as classified in Uniprot-GO (36% of the selected complexes no related to TF activity) and experimental method used to determine the structure retrieved from the PDB.

²⁻³ Details on the protein-DNA binding recognition site. Localization of the DNA-protein interactions in each structure retrieved from the database DNAProDB (<http://dnaprodb.usc.edu/search.html>)[8]. Contacts in the major groove and minor groove are shown by polar contact map (column 3) and cartoon representation of the complex (column 4). The protein-DNA contact areas are coloured according to the protein secondary structure (helix in red, strand in green and loop in blue respectively).



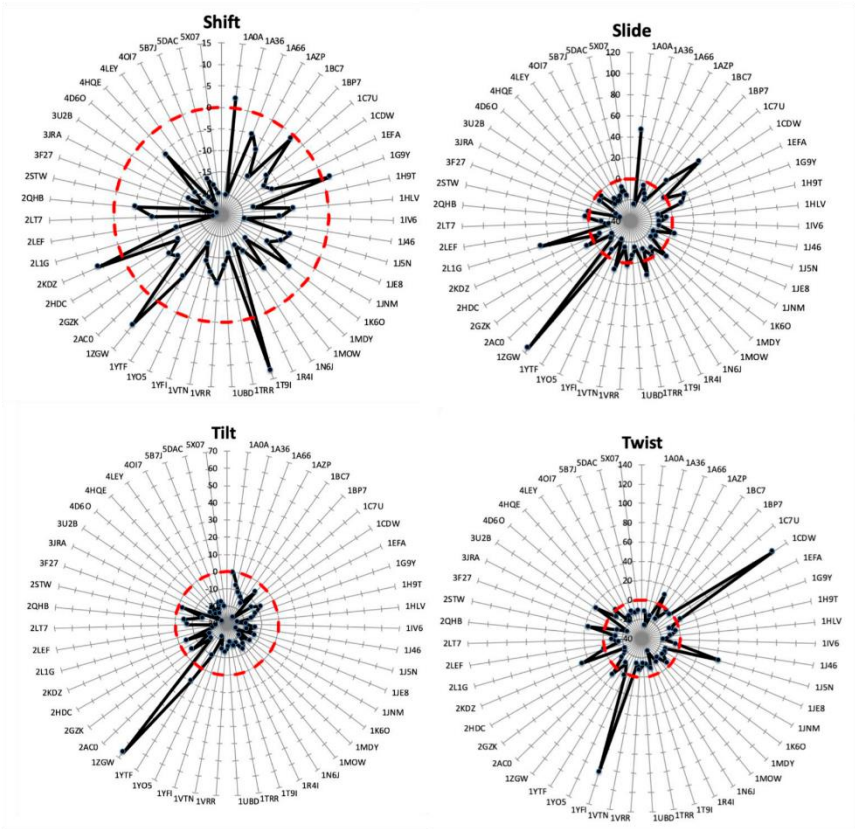
Supplementary Fig. S1. Population of the protein-DNA complexes stored in the Protein Data Bank (PDB), division of the different complexes using our selection depending on DNA structure. After filtering (see Supplementary Methods), 54% of the structures final subset have function related with transcription.



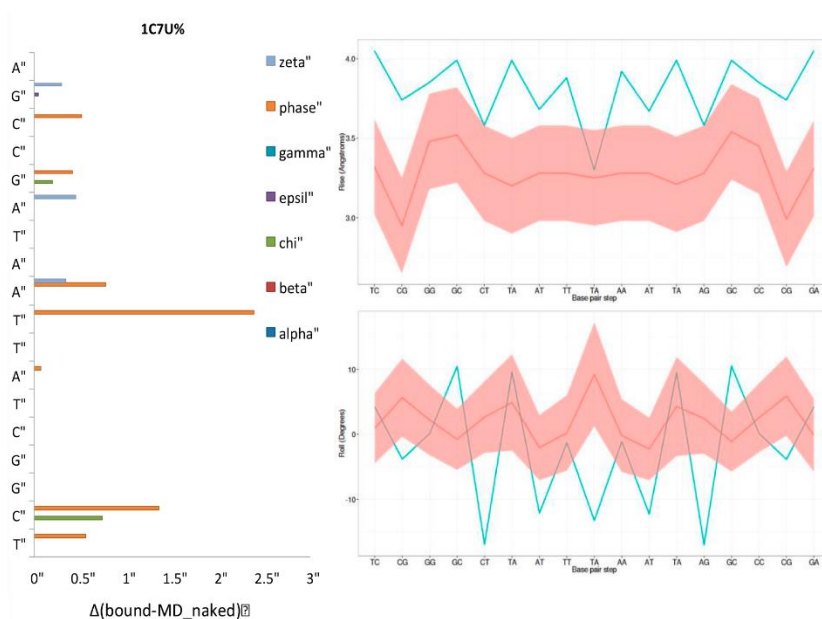


Supplementary Fig. S2. Comparison of base pair step parameters, translational slide and rotational roll, values averaged along the MD simulation for the PDB ids 1IV6 and 1J5N (A and B respectively), starting from the B-ideal DNA (average profile in red with standard deviation contour in pink) and from the experimental

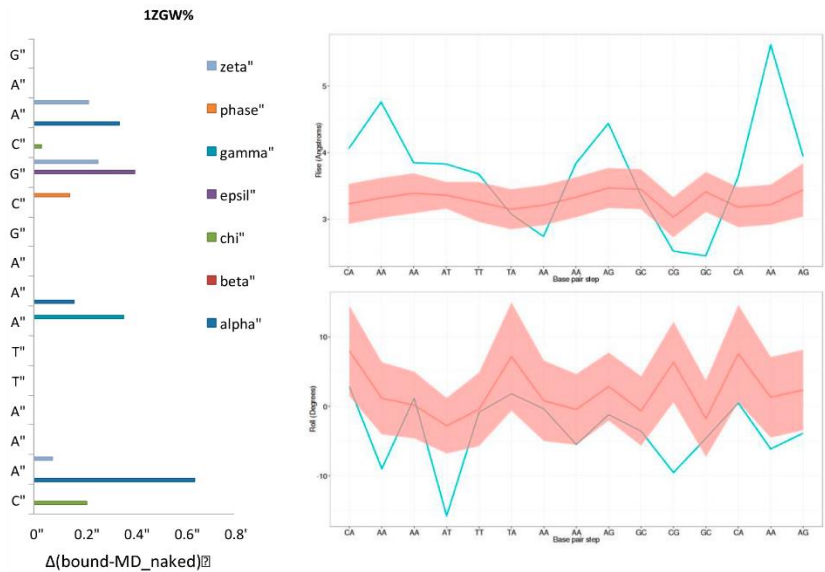
protein-bound DNA structure (average profile in black with standard deviation contour in grey).



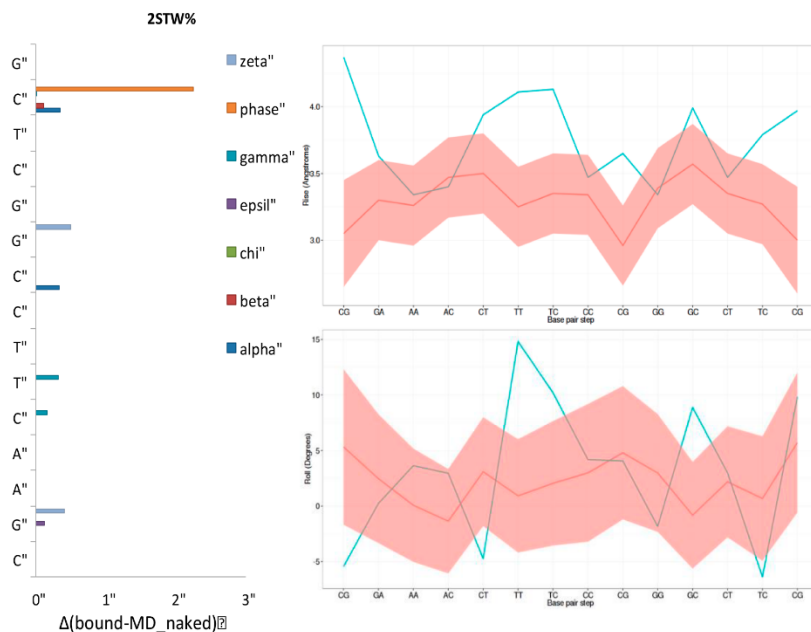
Supplementary Fig. S3. Base pair parameter confidence region profile. For each protein-bound DNA structure identified by their PDB ID, the axis represents the difference between the observed test statistic and the 95% critical value from the F distribution ($F - F_{(1-\alpha; m, n-m)}$). The value for base pair parameter, translation shift and slide and rotational tilt and twist, can be inside (<0 , limit defined by red line) or outside the naked DNA conformational space (>0). See Methods and Supp. Methods for discussion.



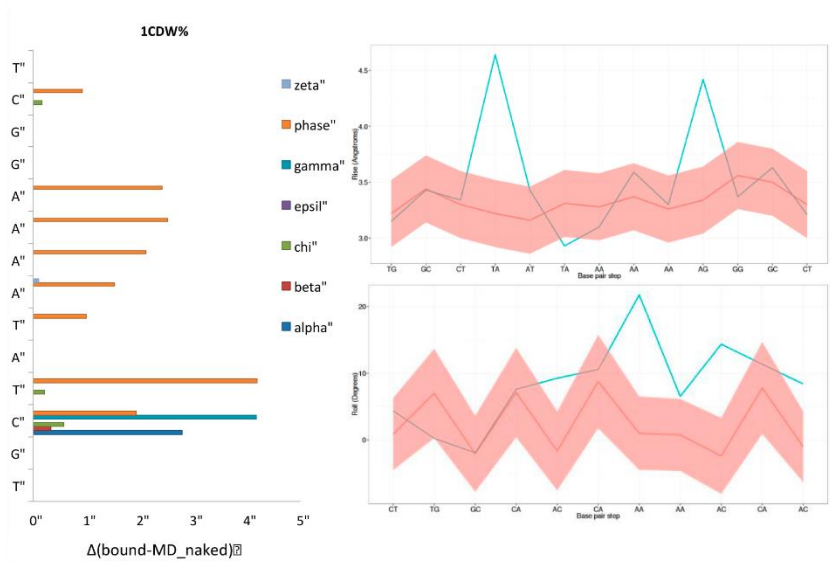
Supplementary Fig. S4. Backbone and base pair parameter analysis for the complex PDB ID 1C7U. On the left the analysis of the backbone angles ($\alpha, \beta, \chi, \epsilon, \gamma, \Pi, \zeta$) is shown. Backbone angles variation has been analyzed using the difference between the experimental and the average MD value plus the standard deviation, divided by the standard deviation along the MD trajectory ($\Delta(\text{bound-MD_naked})$). On the right the comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for each base pair step parameter along some of the sequences found as outlier in Figure 2. Along the full length of the DNA we noticed very extreme values of roll and rise in the protein-bound structure, correlated for some steps with unusual phase angle in the backbone. Some of these distorted base pair steps (TA-AT) are in contact with coil residues (GLY1, GLY101, ARG2, ARG102); while the other steps are only mildly interacting with alpha helices not justifying the high rise distances.



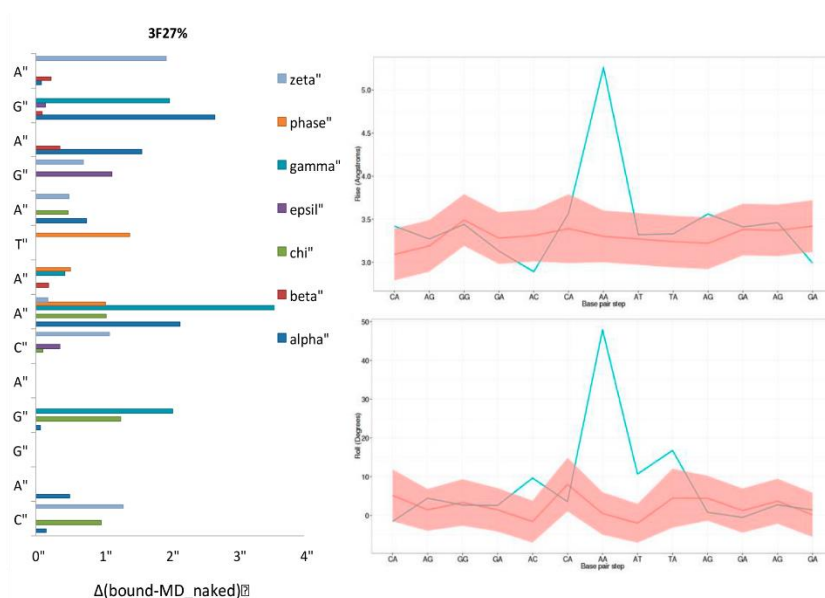
Supplementary Fig. S5. Backbone and base pair parameter analysis for the complex PDB ID 1ZGW. On the left the analysis of the backbone angles ($\alpha, \beta, \chi, \epsilon, \gamma, \Pi, \zeta$) is shown. Backbone angles variation ($\Delta(\text{bound-MD}_{\text{naked}})$) has been analyzed using the difference between the experimental and the average MD value plus the standard deviation, divided by the standard deviation along the MD trajectory. On the right the comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for each base pair step parameter along some of the sequences found as outlier in Figure 2. For this complex we noticed that DNA distortions at the base pair level (at the steps AA, AT, AA respectively) with high absolute values of roll and rise, are correlated with unusual backbone angles mainly in the α, χ and ζ angles. These distortions are in areas where the protein coils (residues THR33-34, ILE36 and ARG45) and alpha helix (PHE114, ARG118, LYS121, PRO128) contact the DNA, even if the mild protein-DNA interactions cannot explain the high base pair deformation.



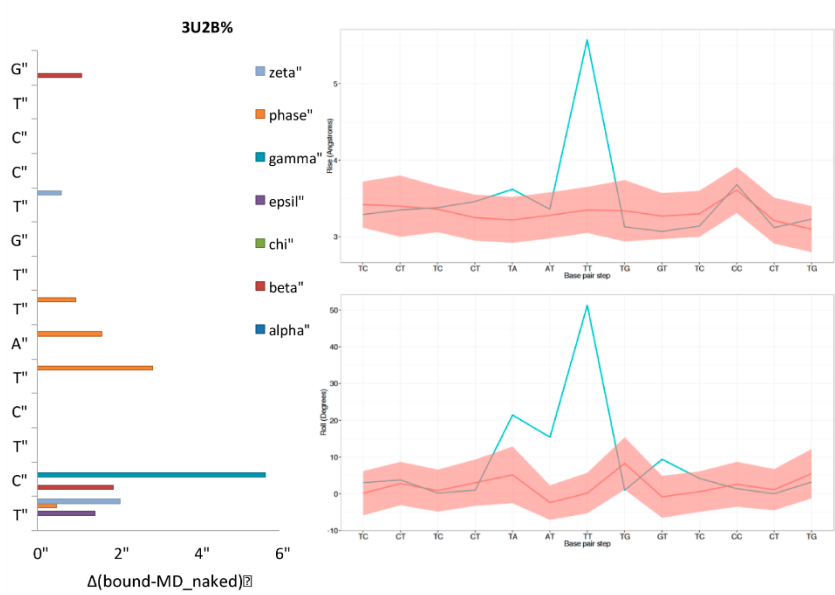
Supplementary Fig. S6. Backbone and base pair parameter analysis for the complex PDB ID 2STW. On the left the analysis of the backbone angles ($\alpha, \beta, \chi, \epsilon, \gamma, \Pi, \zeta$) is shown. Backbone angles variation ($\Delta(\text{bound-MD_naked})$) has been analyzed using the difference between the experimental and the average MD value plus the standard deviation, divided by the standard deviation along the MD trajectory. On the right the comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for each base pair step parameter along some of the sequences found as outlier in Figure 2. The central part of the DNA, in contact with protein helices, is bent and with high value of rise (base pair TT), with correlated distortion in backbone angle χ .



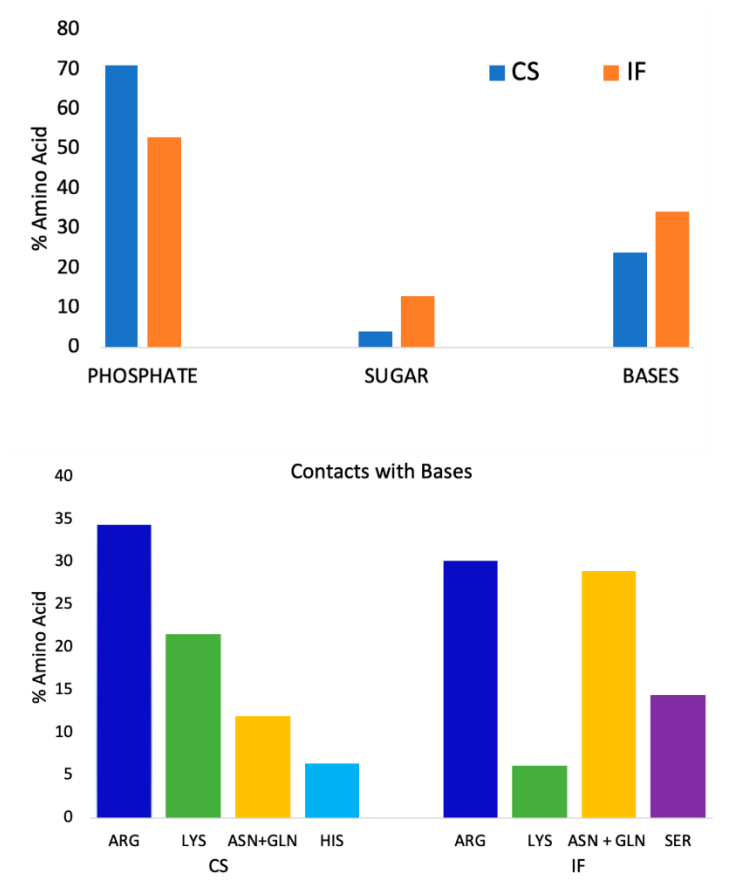
Supplementary Fig. S7. Backbone and base pair parameter analysis for the complex PDB ID 1CDW. On the left the analysis of the backbone angles ($\alpha, \beta, \gamma, \epsilon, \gamma, \Pi, \zeta$) is shown. Backbone angles variation ($\Delta(\text{bound-MD_naked})$) has been analyzed using the difference between the experimental and the average MD value plus the standard deviation, divided by the standard deviation along the MD trajectory. On the right the comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for each base pair step parameter along some of the sequences found as outlier in Figure 2. The distortion given by the contact of the protein β - β residues (PHE283, LEU299, THR309, VAL255, ASN253, VAL165, PHE210 and LEU208) at the base level, high roll and rise values at steps TA, AA, AG respectively, is correlated with deformation of the phase angle in the backbone.



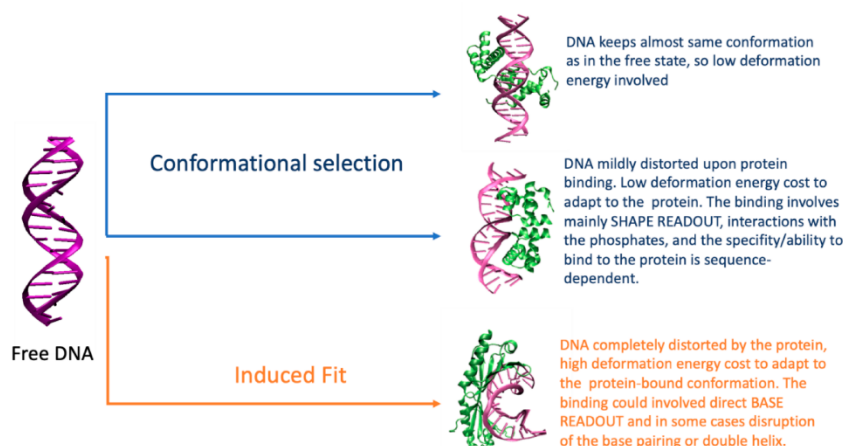
Supplementary Fig. S8. Backbone and base pair parameter analysis for the complexes that showed main/critical distortions compared to the unbound DNA structures PDB ID 3F27. On the left the analysis of the backbone angles ($\alpha, \beta, \gamma, \epsilon, \chi, \pi$ and ζ) is shown. Backbone angles variation ($\Delta(\text{bound-MD_naked})$) has been analyzed using the difference between the experimental and the average MD value plus the standard deviation, divided by the standard deviation along the MD trajectory. On the right the comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for each base pair step parameter along some of the sequences found as outlier in Figure 2. For this complex we noticed that where the protein helices (residues PHE75, MET76, SER99, ASN95, ARG83) contact the DNA, central base pairs AATA, the DNA is highly distorted with high values of roll and rise, that are correlated with a distortion in the backbone angles, in particular in the γ angle where the DNA is bent/kinked.



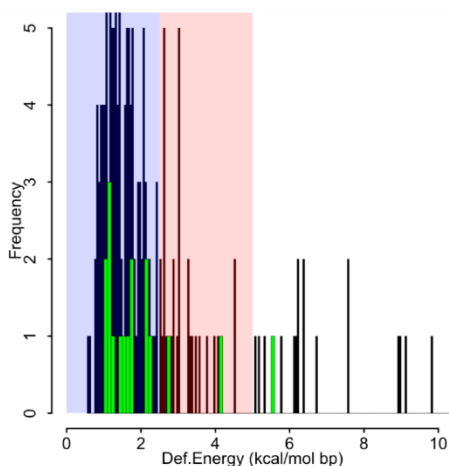
Supplementary Fig. S9. Backbone and base pair parameter analysis for the complexes that showed main/critical distortions compared to the unbound DNA structures PDB ID 3U2B. On the left the analysis of the backbone angles ($\alpha, \beta, \gamma, \epsilon, \zeta$ and χ) is shown. Backbone angles variation ($\Delta(\text{bound-MD_naked})$) has been analyzed using the difference between the experimental and the average MD value plus the standard deviation, divided by the standard deviation along the MD trajectory. On the right the comparison between the experimental (blue) and MD values (red with standard deviation contour in pink) for each base pair step parameter along some of the sequences found as outlier in Figure 2. For this complex we noticed that where the protein helices (residues PHE10, MET11, SER34, ALA31) contact the DNA, central step TT, the DNA is highly distorted with high values of roll and rise, that are correlated with a distortion in the backbone in the phase angle.



Supplementary Fig. S10. Top panel: Percentage of amino acid interacting with the phosphate, sugar or bases of the DNA (distance less than 5 Å). Data for the 2 recognition group identified from our full set of protein-DNA complexes, conformational selection group with deformation energy < 2.5 kcal/mol bp (blue) and induced fit group (>5kcal/mol, red). Bottom panel: percentage of the type of amino acids interacting with the bases of the DNA in the two group, conformational selection (CS) and induced fit (IF).



Supplementary Fig. S11. Scheme of the different recognition modes, conformational selection and induced fit, with the relative characteristics.



Supplementary Fig. S12. Frequency of the deformation energy cost (kcal/mol·bp) required moving from the unbound to the bound conformation in the helical space for all the DNA-protein interactome. In black bars represent the deformation energy for the structure in the selected dataset, the green bars the values for the structure hand-curated to remove modified bases and unpaired bases. In this distribution also the hand-curated structure fall predominantly within the area with energy < 2.5 kcal/mol·bp (blue area).

SUPPLEMENTARY REFERENCES

- [1] B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A.I. Petrov, B. Sweeney, C.L. Zirbel, N.B. Leontis, H.M. Berman, The Nucleic Acid Database: new features and capabilities, *Nucleic Acids Res.* 42 (2014) D114–D122. doi:10.1093/nar/gkt980.
- [2] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* 81 (1984) 3684–3690. doi:10.1063/1.448118.
- [3] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems, *J. Chem. Phys.* 98 (1993) 10089–10092. doi:10.1063/1.464397.
- [4] J.-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes, *J. Comput. Phys.* 23 (1977) 321–341. <http://physics.ujep.cz/~mlisal/md/shake.pdf> (accessed May 22, 2017).
- [5] I. Ivani, P.D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J.L. Gelpí, C. González, M. Vendruscolo, C.A. Loughton, S.A. Harris, D.A. Case, M. Orozco, Parmbsc1: a refined force field for DNA simulations, *Nat. Methods.* 13 (2015) 55–8. doi:10.1038/nmeth.3658.
- [6] P.D. Dans, I. Ivani, A. Hospital, G. Portella, C. González, M. Orozco, How accurate are accurate force-fields for B-DNA?, *Nucleic Acids Res.* 44 (2017) gkw1355. doi:10.1093/nar/gkw1355.
- [7] P.D. Dans, L. Danilâne, I. Ivani, T. Dršata, F. Lankaš, A. Hospital, J. Walther, R.I. Pujagut, F. Battistini, J.L. Gelpí, R. Lavery, M. Orozco, Long-timescale dynamics of the Drew–Dickerson dodecamer, *Nucleic Acids Res.* 44 (2016) 4052–4066. doi:10.1093/nar/gkw264.
- [8] J. Li, J.M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez, R. Rohs, Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding, *Nucleic Acids Res.* 45 (2017) 12877–12887. doi:10.1093/nar/gkx1145.

III. Determinants of nucleosome architecture in yeast

SUPPLEMENTARY DATA

The interplay between periodicity, DNA physical properties and effector binding define nucleosome architecture in yeast

Diana Buitrago^{1,&}, Mireia Labrador^{1,&}, Pau De Jorge¹, Federica Battistini¹, Isabelle Brun Heath¹ and Modesto Orozco^{1,2*}

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain; ² Departament de Bioquímica i Biomedicina, Universitat de Barcelona, Barcelona, Spain

& These authors contributed equally to this work

* Correspondence to M.Orozco: modesto.orozco@irbbarcelona.org

Contents

Supplementary Materials and Methods	1
Supplementary Figures	4
Supplementary Tables	10

Supplementary Materials and Methods

Mutant strains generation

Delitto perfetto (Storici and Resnick 2006) : Briefly, the first step consisted on inserting the URA3 gene in the gene of interest. This involved amplifying the URA3 gene from the YDp-U vector. This cassette was then integrated by homologous recombination into the target gene and recombinants were selected on selective medium without uracil. The second step consisted on the insertion of the seq81 DNA sequence into the gene of interest in replacement of the URA3 gene. It required the cloning the seq81 DNA sequence into the pCR™-Blunt II-TOPO™ Vector, the amplification of this fragment with the specific oligonucleotides so that the URA cassette is lost when the seq81 is integrated. Transformed colonies were selected on YPD medium supplemented with 5-FOA and the insertion were verified by PCR and by standard DNA sequencing.

RNA extraction and RT-qPCR. Exponential cultures were arrested at late G1 by alpha-factor. RNA was obtained from 20ml yeast cultures (OD_{600} 0.8) using the hot-phenol method.: cells were centrifuged, and pellets were resuspended in 400 μ l of AE buffer (50 mM NaOAc and 10 mM EDTA) with 1% SDS. In the same tube, approximately 0.5 ml of glass beads and 410 μ l (1 vol) of acid phenol were added. Cells were then vortexed and incubated for 10' at 65°C followed by 5' in ice. Cells were then centrifuged to collect the supernatant that was treated with one volume of phenol/chloroform followed by ethanol precipitation. The obtained RNAs were resuspended in 40 μ l of RNase-free MilliQ water and finally quantified using the Nanodrop and treated by DNase I (1u/ μ g of RNA).

Nucleosome mapping

Semi-intact yeast cells. For nucleosome preparation, semi-intact yeast cells were prepared as described in (Schlenstedt et al. 1993) : a culture of 50 ml of yeast cells was grown at 30°C until OD_{600} 0.8–0.9 and arrested at late G1 with alpha-factor. When applicable, transcription was inhibited by adding 10-phenanthroline (100 μ g/ml) for 30 minutes at 30°C with shaking. Cells were then harvested at 700g for 7'. Supernatants were discarded and the pellets were resuspended in 5 ml of 100 mM PIPES pH 9.4 containing 10 mM DTT and incubated for 10' at 30°C with continuous shaking. Cells were then centrifuged at 1000g for 10', resuspended in 1.2 ml of YEP 0.2% glucose buffer (1% bacto-yeast extract, 1% bacto peptone, 0.5% NaCl (w/V), 50 mM KH_2PO_4 , 0.2% glucose (w/V), 0.6 M sorbitol) and 30 U/ml zymolase (G-Biosciences, 786-036) was added. After 30' of incubation at 30 °C with continuous shaking, 80% of the cells were spheroplasts. The spheroplast were then centrifuged at 1000 g for 5' resuspended in 8 ml of YEP 1% glucose buffer (1% bacto-yeast extract, 1% bacto peptone, 0.5% NaCl (w/V), 1% glucose (w/V), 0.7 M sorbitol) and incubated for 20' at 30°C. Samples were then spun down at 1000 g for 5' and washed twice with 1 ml of permeabilization buffer (20 mM PIPES-KOH pH 6.8, 150 mM KAc, 2 mM MgAc, 0.4 M sorbitol) plus 10% of DMSO. Finally samples were resuspended in 0.25 ml of permeabilization buffer, frozen in liquid nitrogen and stored at -80°C.

MNase digestion: To determine the MNase digestion conditions that produce >80% mononucleosomal DNA fragments, we performed a digestion optimization using different MNase digestion times with a small amount of semi-intact cells from every batch preparation. 50 μ l of semi-intact cells were thawed on ice with 503 μ l of EcoRI buffer (50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 100 mM NaCl, 0.02% Triton X-100 10%, 0.1 mg/mL BSA), 3 mM $CaCl_2$ and 0.5 mM of MNase (Sigma, N-3755). The digestion reactions were incubated at 37°C and stopped at different time points (0', 5', 10', 15', 20', 30 and 40') by taking out 80 μ l of the mix and adding 70 μ l of stop solution (100 mM EDTA). DNA was then treated with 50 μ g/ml RNase A, incubated 4 hours at 37°C and digested with 0.22 mg/ml of proteinase K plus 1.5% SDS for at least 12 hours at 65°C. DNA samples were purified twice by phenol/chloroform/isoamyl extraction and ethanol precipitation. Finally, samples were resuspended in 10 μ l of water and

examined in 2% agarose gels. Once the samples were ready, the percentage of mononucleosomal DNA fragments were examined by 2% agarose gels.

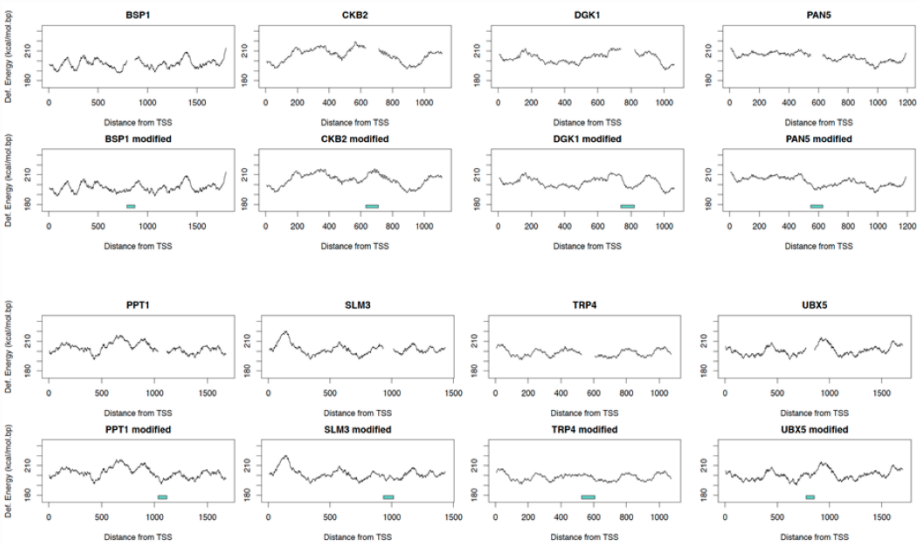
TFBS density

Global transcription factor binding site (TFBS) affinity was estimated from TRANSFAC position weight matrices (PWMs). Binding site predictions from every PWM were computed for yeast genome, using R/Bioconductor Biostrings library with default parameters. A global TFBS density was computed pooling all the predicted sites and computing their coverage.

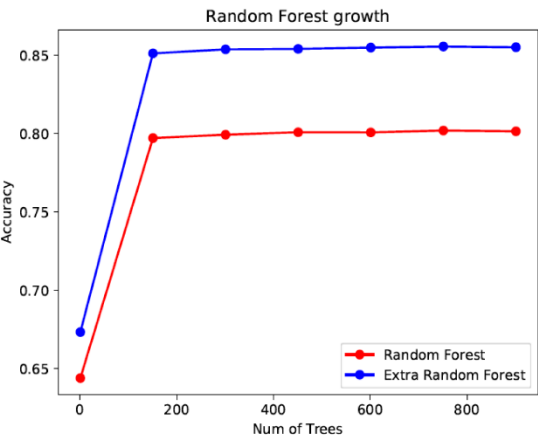
Schlenstedt G, Hurt E, Doye V, Silver PA. 1993. Reconstitution of nuclear protein transport with semi-intact yeast cells. *J Cell Biol* **123**: 785-798.

Storici F, Resnick MA. 2006. The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods in enzymology* **409**: 329-345.

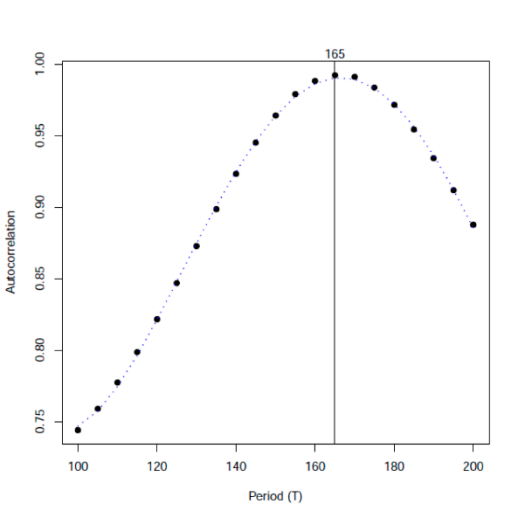
Supplementary Figures



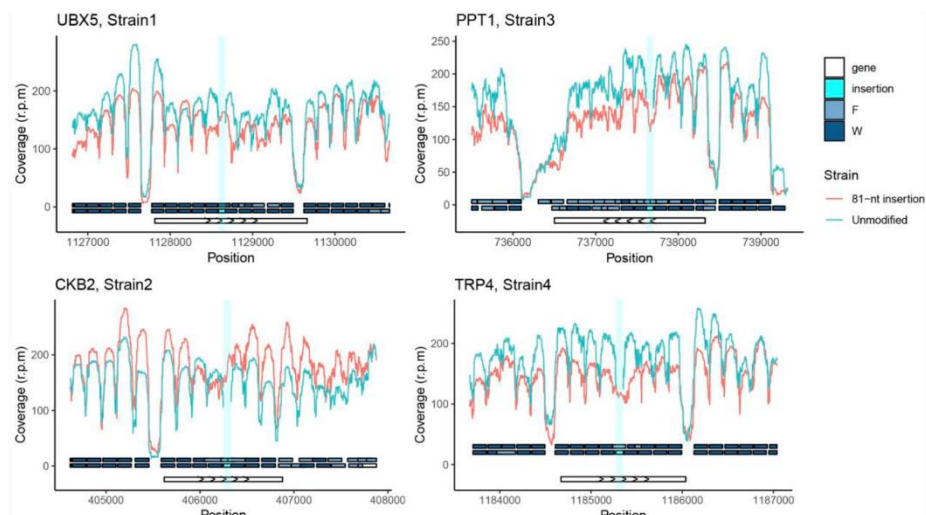
Suppl. Figure S1: Deformation energy for the 8 selected genes without (top panels), or with (lower panels) the 81-nt insert (represented as a box colored in cyan).



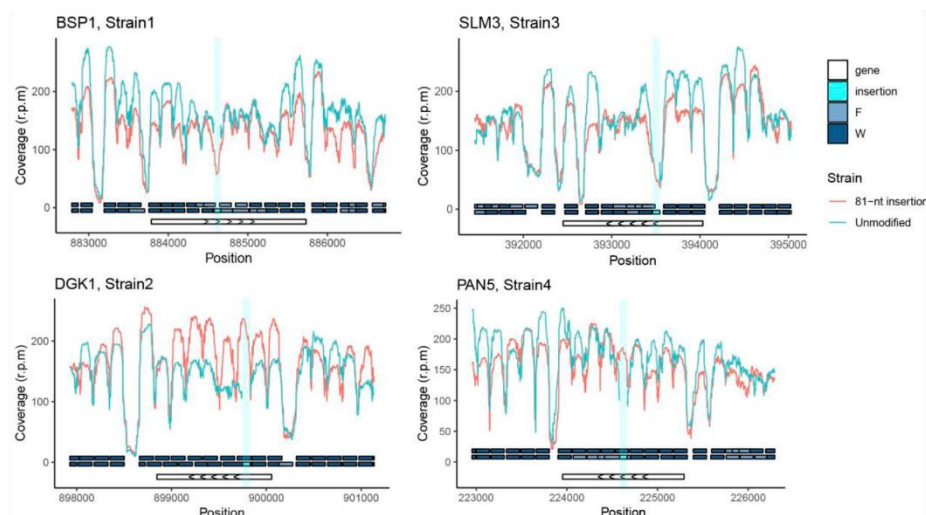
Suppl. Figure S2. Accuracy of Random Forest (red) and Extra-Trees (blue) classifiers in the validation set, for different numbers of trees.



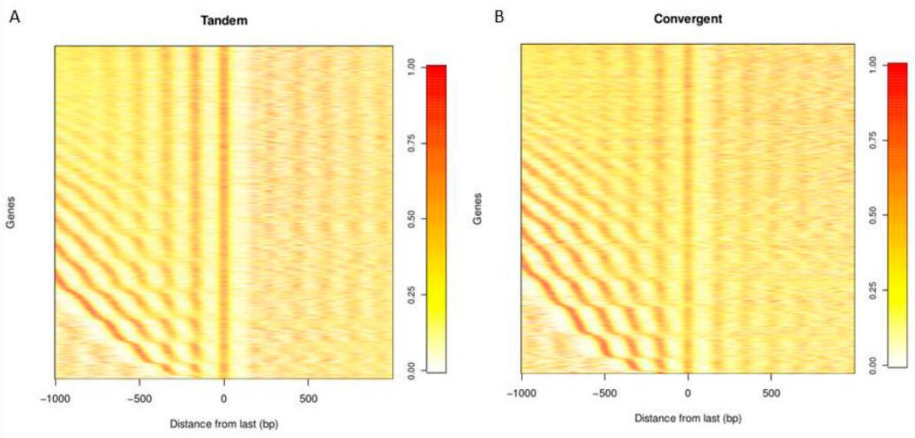
Suppl. Figure S3. Autocorrelation coefficient for the nucleosome coverage signal in different potential periods



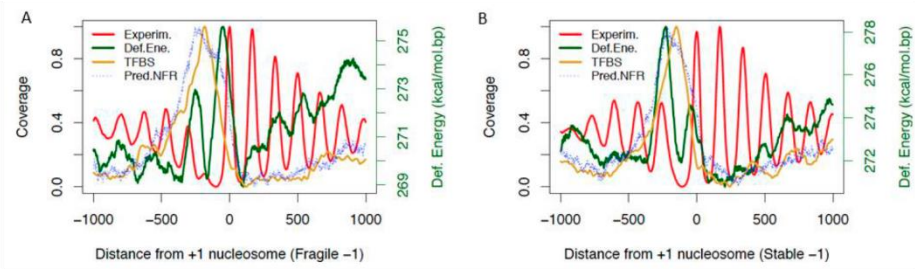
Suppl. Figure S4. Nucleosome coverage of the four phased genes in the unmodified strain (blue) and their nucleosome coverage in the strain with the 81-nt insertion (red). Nucleosome positions in the original and modified strains are shown colored by their class: Well-positioned (W) and Fuzzy (F). The gene body from TSS to TTS is depicted as a white box.



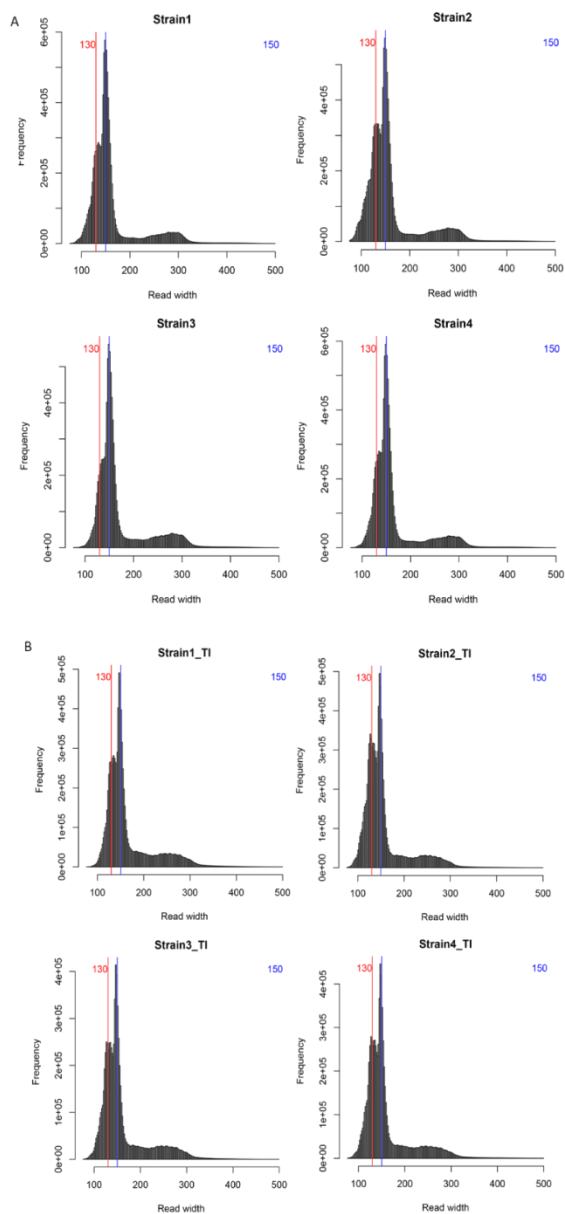
Suppl. Figure S5. Nucleosome coverage of the four not-phased genes in the unmodified strain (blue) and their nucleosome coverage in the strain with the 81-nt insertion (red). Nucleosome positions in the original and modified strains are shown colored by their class: Well-positioned (W) and Fuzzy (F). The gene body from TSS to TTS is depicted as a white box.



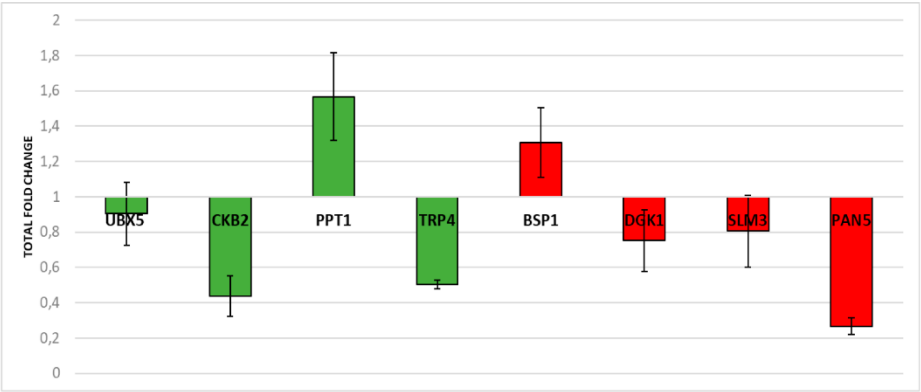
Suppl. Figure S6. Nucleosome coverage of (A) tandem and (B) convergent genes, centered at $-last$ nucleosome. Genes are sorted by the distance between $+1$ and $-last$ nucleosomes.



Suppl. Figure S7. Average nucleosome coverage (red), TFBS density (yellow), deformation energy (green) and NFR prediction (blue) around $+1$ nucleosome, for genes with -1 nucleosome classified fragile (left panel) or stable (right panel). Fragile and stable classification was obtained from Kubik et al. (2015).



Suppl. Figure S8. Length of the sequenced fragments in the MNase-seq experiments of samples (A) with transcription and (B) inhibited by 1,10 phenantroline.



Suppl. Figure S9. Gene Expression comparison (qPCR) for each gene with or without the 81-nt sequence insertion.

Supplementary Tables

Suppl. Table S1. Genes modified with the 81-nt sequence in each strain

Gene	Strain	Strand	Start	End	Chromosome	Insert position	Phasing
UBX5	1	+	1127872	1129374	chrIV	1128586	Phased
CKB2	2	+	405768	406544	chrXV	406248	Phased
PPT1	3	-	736662	738203	chrVII	737615	Phased
TRP4	4	+	1184747	1185889	chrIV	1185196	Phased
BSP1	1	+	883828	885558	chrXVI	884578	Control (not-phased)
DGK1	2	-	899056	899928	chrXV	899667	Control (not-phased)
SLM3	3	-	392659	393912	chrIV	393462	Control (not-phased)
PAN5	4	-	224030	225169	chrVIII	224581	Control (not-phased)

Suppl. Table S2. Sequences of primers used in the gene expression analysis by qPCR, and the amplicon size (in base pairs) obtained using each forward (F) and reverse (R) pair of oligonucleotides.

Name	Primer sequence (5'-3')	Amplicon size (bp)
UBX5_F	GACGACGACGAATATGAG	163
UBX5_R	CGAGTTTGGACATGATTG	
CKB2_F	GAGAATATGACACATGCC	130
CKB2_R	CCGCCTCTTTGACTTAG	
PPT1_F	CAATGATCCGGCTGCTAC	171
PPT1_R	GGACCTTCATAATTGGCT	
TRP4_F	GTAGGTACTGGTGTGAC	117
TRP4_R	GGATGTAGAAGCTTTACC	
BSP1_F	GGAAGAGCCGATATACCC	147
BSP1_R	CGGACTTTCTGTAACCTC	
DGK1_F	CCATTGCCCTTCCAAATA	182
DGK1_R	GCCGAACCATTGATGAGA	
SLM3_F	GATTGGAGAGATGTGAAC	123
SLM3_R	CGACCCTTCACTGTAGCC	
PAN5_F	GGGTACCGTTTTGGCAGT	178
PAN5_R	GCATTCGCATTCTCCAC	
PMA1_F	TTTGCCAGCTGTCGTTACCA	240
PMA1_R	TTCTTCTTTCTGGAAGCAGC	
ACT1_F	GGTTGCTGCTTTGGTTATTGATAAC	271
ACT1_R	CAATTCGTTGTAGAAGGTATGATGCC	
RPA135_F	GAACAATGGCGAGGAGAAC	141
RPA135_R	CCACCTATTTTCATCGGATTC	
NMD3_F	GGGTTGGATTTCTTCTATGC	164
NMD3_R	GGAACAATCTCGACAGAATATG	

Suppl. Table S3. Phase score (DFI) and autocorrelation (R) in the unmodified strain and in the strain with the 81-nt insertion in the selected genes. Genes UBX5, CKB2, PPT1 and TRP4 are phased, and genes BSP1, DGK1, SLM3 and PAN5 are not-phased in the unmodified strain.

Gene	Unmodified strain		Strain with the 81-nt insert	
	DFI	R	DFI	R
UBX5	7	0.79424	79	0.76687
CKB2	1	0.81862	79	0.68771
PPT1	13	0.88553	71	0.82500
TRP4	6	0.80702	80	0.73771
BSP1	77	0.73841	7	0.77164
DGK1	46	0.69007	31	0.81921
SLM3	37	0.61947	42	0.60019
PAN5	44	0.73341	46	0.81708

IV. Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning

SUPPLEMENTARY DATA

Nucleosome Dynamics: A new tool for the dynamic analysis of nucleosome positioning

Diana Buitrago^{1&}, Laia Codó^{2&}, Ricard Illa¹, Pau de Jorge¹, Federica Battistini¹, Oscar Flores¹, Genis Bayarri¹, Romina Royo², Marc Del Pino², Simon Heath³, Adam Hospital¹, Josep Lluís Gelpí^{2,4}, Isabelle Brun Heath¹ and Modesto Orozco^{1,4*}

¹ Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri i Reixach 10. Barcelona 08028. Spain.

² Barcelona Supercomputing Center (BSC). Jordi Girona 31. Barcelona 08028. Spain.

³ CNAG. Centro Nacional de Análisis Genómico.

⁴ Departament de Bioquímica i Biomedicina. Facultat de Biologia. Universitat de Barcelona. Avgda Diagonal 647. Barcelona 08028. Spain.

& Equally contributing authors

Contents

Supplementary Methods	1
Supplementary Figures	3
Supplementary Tables	12

Supplementary Methods

Occupancy and fuzziness measurements

In the simulated data, nucleosome occupancy was defined as the number of cells that contain a read corresponding to the each nucleosome, whereas fuzziness is described by the standard deviation of the centers from all reads overlapping the synthetic nucleosome position.

We executed nucleR and DANPOS on the generated *in silico* nucleosome maps to compare scores of occupancy and fuzziness of the nucleosome positioning algorithms. From nucleR, `score_height` (logit scale) determines the occupancy and `score_width` the fuzziness. From DANPOS we use `smt_value` (occupancy) and `fuzziness_score`.

We compared the true values of occupancy and fuzziness to the values obtained from nucleR and DANPOS and computed for each measurement the R^2 of the linear relation between them.

Data resources

MNase-seq datasets reported as pilot cases in this study were obtained from the ENA-SRA website (<http://www.ebi.ac.uk/ena>) and the GEO repository under accession numbers: PRJEB6970 for the cell cycle data, GSE77631 for the yeast metabolic cycle and GSE13622 for the nucleosome maps from yeast cultivated in glucose, galactose and ethanol media. Raw data were mapped using Bowtie aligner (Langmead et al. 2009) to the SacCer3 reference genome (downloaded from <http://hgdownload-test.cse.ucsc.edu/goldenPath/sacCer3/bigZips>).

Yeast transcription start sites (TSS) were compiled from Pelechano et al. (2013), Miura et al. (2006), Yassour et al. (2009) and Nagalakshmi et al. (2008).

GO enrichment analysis

The genes with mapped nucleosome architecture changes (inclusions, evictions and shifts) around their promoter (-350 base pairs upstream and +50 base pairs downstream from the TSS) were summarized in terms of the biological functions they relate to. We performed Gene Ontology (GO) enrichment analysis using GOSTATS R package (Falcon and Gentleman 2007), which employs a Hypergeometric test to find over-represented GO terms in the list of genes, adjusting for multiple testing with Benjamini–Hochberg correction.

Falcon S, Gentleman R (2007). "Using GOSTATS to test gene lists for GO term association." *Bioinformatics*, **23**(2), 257-8.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. *Genome Biol.* **10**, R25 (2009).

Miura, F. et al. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 17846–17851 (2006).

Nagalakshmi, U. et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344–1349 (2008).

Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**, 127–131 (2013).

Yassour, M. et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 3264–3269 (2009).

Supplementary Figures

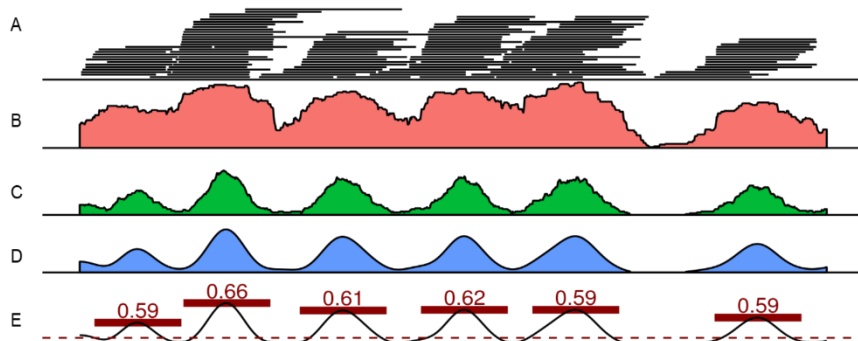


Figure S1. NucleR pipeline to call nucleosomes from MNase-seq experimental data. (A) Reads mapped to reference genome. (B) Coverage of nucleosomal reads per base pair. (C) Coverage of reads trimmed around their centre. (D) Signal smoothing with Fast Fourier Transform. (E) Peak calling and nucleosome scoring.

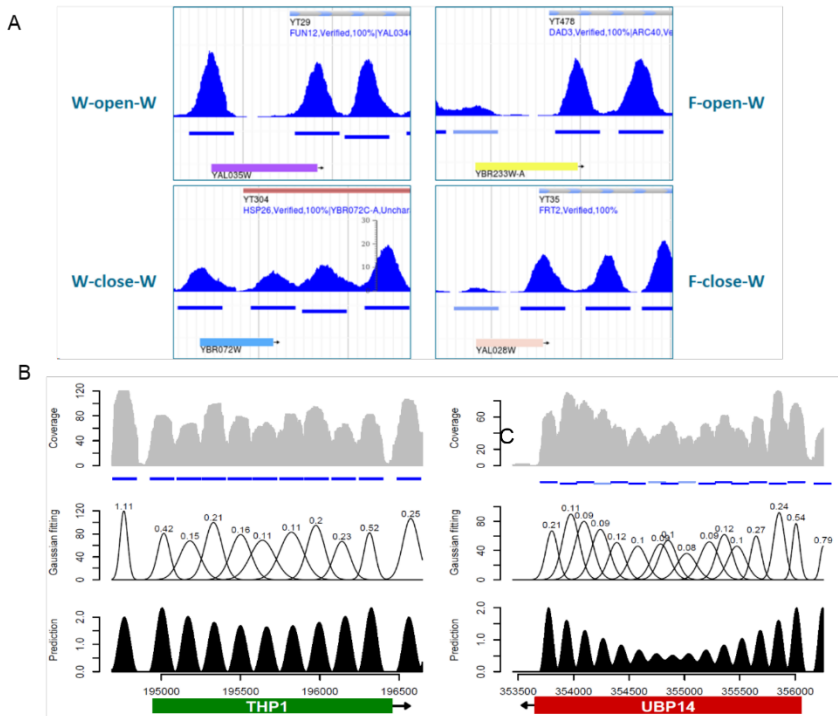


Figure S2. Further analyses for nucleosome calls in an MNase-seq experiment: (A) Promoter classification according to surrounding nucleosomes. Fuzzy nucleosomes (F) are characterized by low and broad peaks, while well-positioned nucleosomes (W) correspond to high and narrow peaks. (B) A phased-gene and (C) a non-phased gene: nucleosome coverage in G1-synchronized cells is shown in grey and nucleosomes detected by nucleR in blue. In the second panel (black curves), a normal distribution is fitted read coverage of each nucleosome, and the estimated standard deviation is used to compute nucleosome stiffness (shown on top of each nucleosome normal curve). The third panel (black coverage) shows reconstituted nucleosome coverage from two signals in opposing directions: downstream from +1 nucleosome and upstream from -last nucleosome. “Phased” genes coloured in green, “antiphased” genes in red.

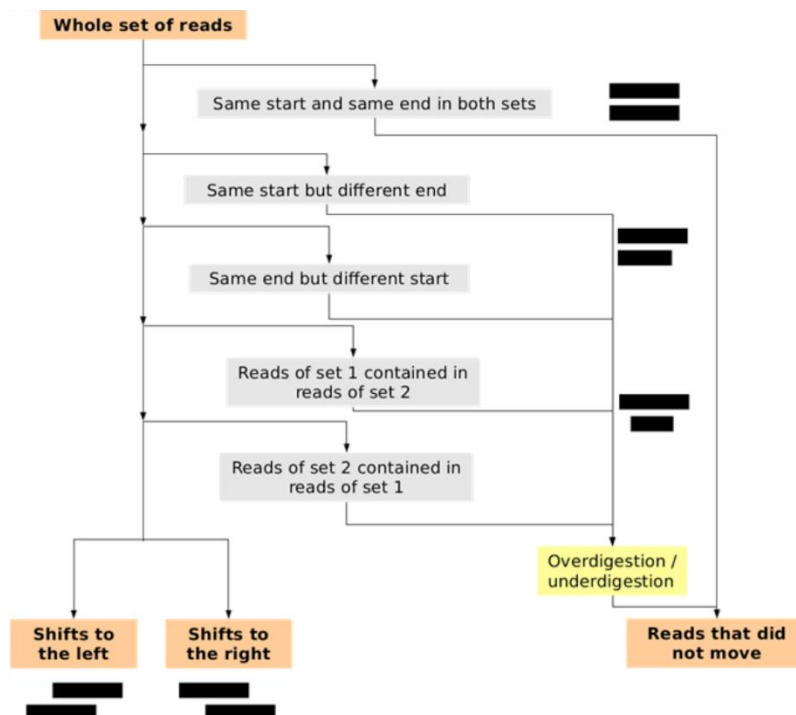


Figure S3. NucDyn pipeline to discard reads which do not change between two MNase-seq experiments. Mapped fragments that are equal in the two experiments, have the same start or end, or are contained within a fragment from the other experiment are discarded.

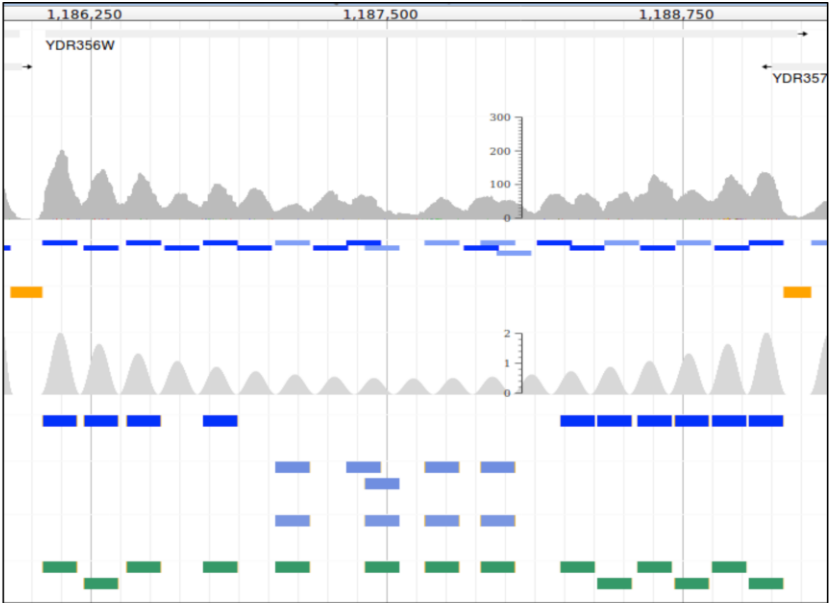


Figure S4. Synthetic nucleosome maps are generated based on periodic signals from the NFRs (yellow). Dark grey coverage is obtained from experimental MNase data (G1 synchronized cells) and the corresponding nucleosome calls obtained with nucleR are shown below. Light grey coverage shows the periodic signals used to position synthetic nucleosomes (the height of each peak is the probability of a nucleosome to be present in a family). Below them, dark blue boxes represent synthetic nucleosome fragments for W nucleosomes, and light blue correspond to F nucleosomes in a single family. In case of overlap between generated nucleosomes in a single family, one of the two reads is randomly removed. The final synthetic nucleosome positions from a single family are shown in green.

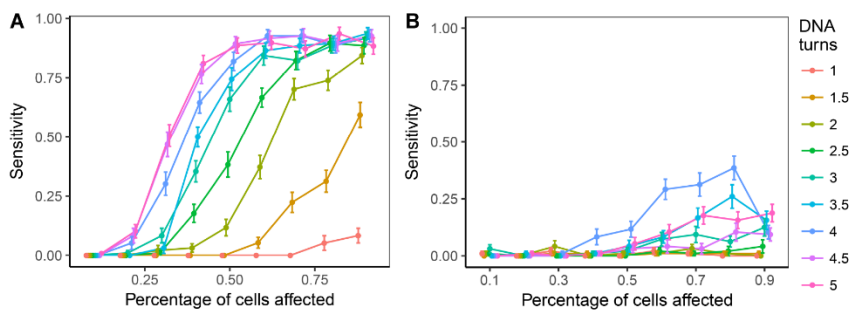


Figure S5. Sensitivity of the SHIFT prediction for (A) NucDyn and (B) DANPOS, computed on synthetic nucleosome maps. Shifts were introduced displacing reads from 1 to 5 DNA turns and modifying different percentages of the nucleosome families (10%, 20%, ..., 90%). A shift was detected when DANPOS FDR < 0.01 and the distance between nucleosomes ($\text{treat2control_dis} - 10$) is larger than the corresponding displacement. NucDyn was run with default parameters.

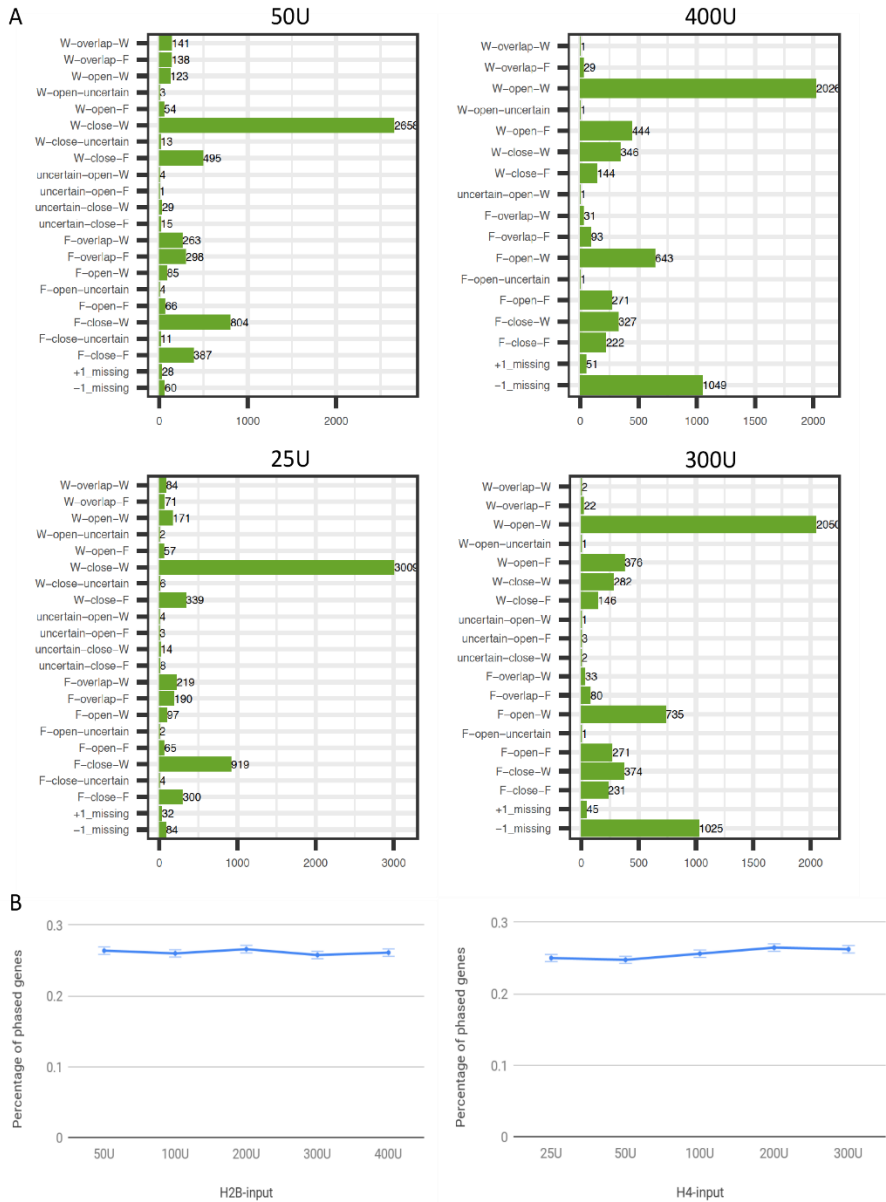


Figure S6: Comparison of genome wide nucleosome positioning at promoters between the different levels of MNase digestion. MNase-seq data obtained from Chereji et al. 2017. (A) TSS classification for H2B-inputs (top panel, 50U and 400U) and H4-inputs (lower panel, 25U and 300U) under different MNase concentrations. (B) Percentage of phased genes under different MNase levels.

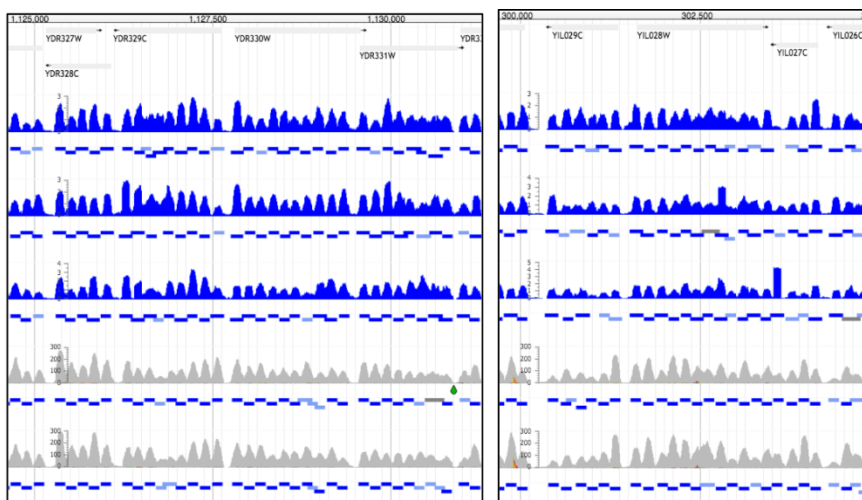


Figure S7: Comparison in two regions of the yeast genome of the chemical cleavage (blue coverage tracks) nucleosome positioning data for three replicas from Chereji et al. (2018) and MNase-seq (grey coverage tracks) in cell-cycle synchronized cells in G1 (top) and S (bottom track) phase from Deniz et al. (2016). Nucleosome positions obtained with nucleR are shown as blue boxes coloured by their positioning (W: dark, F: light).

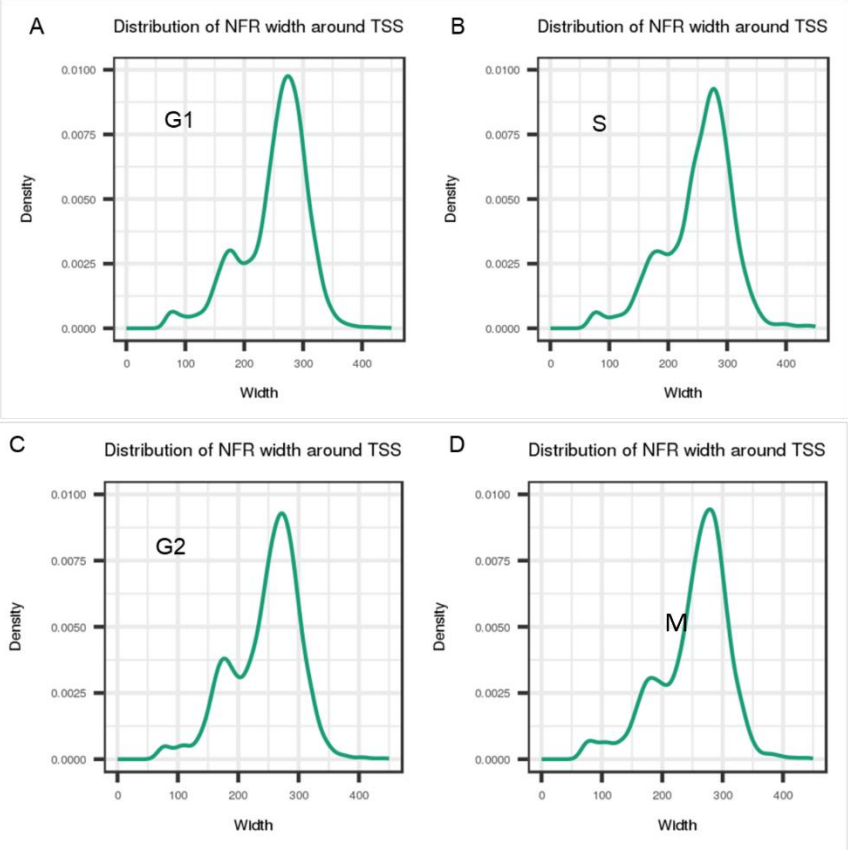


Figure S8: Comparison of genome wide nucleosome positioning at promoters between the different phases of the cell cycle. NFR width in (A) G1 phase, (B) S phase, (C) G2 phase and (D) M phase.

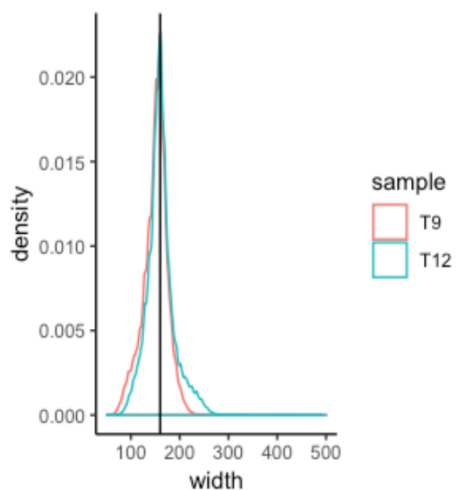


Figure S9: Width of the MNase-seq mapped fragments from Nocetti and Whitehose (2016), for time points 9 and 12 of the yeast metabolic cycle.

Supplementary Tables

Table S1. List of analyses available in Nucleosome Dynamics and description of their parameters.

nucleR	
Parameter	Description
input:	Input BAM file (RData format)
output:	Nucleosome calls in GFF format. Annotations: Score_weight (0-1), Score_height (0-1), class (W, F, uncertain)
type	Type of sequence data: single paired
minoverlap:	Minimum number of overlapping base pairs in two nucleosome calls for them to be merged into one (bp). Optional, default 80.
width:	Width given to nucleosome calls previous to merging (bp). Optional. Default 147.
dyad_length:	Number of bases around the dyad of the nucleosome calls to be used for nucleosome scoring (bp). Optional, default 50.
hthresh:	Height threshold (between 0 and 1) to classify (in combination with width threshold) a nucleosome as either fuzzy or well-positioned according to the number of reads in the dyad of the nucleosome call. Nucleosomes below this value (that is, nucleosomes with low coverage) will be defined as fuzzy. Optional, default 0.4.
wthresh:	Width threshold (between 0 and 1) to classify (in combination with height threshold) a nucleosome as either fuzzy or well-positioned according to the dispersion of the reads around the dyad. Nucleosomes below this value (that is, nucleosome calls not sharp enough) will be defined as fuzzy. Optional, default 0.6
pcKeepComp	Parameter used in the coverage smoothing when Fourier transformation is applied. Number of components to select with respect to the total size of the sample. Allowed values are numeric (in range 0:1) for manual setting, or 'auto' for automatic detection. Optional, default 0.02.
fdrOverAmp	Threshold to filter over-amplified reads, as defined in filterDuplReads function of htSetqTools R package. Optional, default 0.05.
components	Number of negative binomials that will be used to filter duplicated reads, as defined in filterDuplReads function of htSetqTools R package. Optional, default 1.
fragmentLen	Maximum fragment length allowed (bp). Optional, default 170.
trim	Number of bases around the center of each fragment (bp) to use for peak calling. Optional, default 50.
threshold	Defines what threshold should be used to filter out non-significant nucleosome calls. If set to TRUE, the percentage value is considered (--thresholdPercentage). If set to FALSE, the absolute value (--thresholdValue) is used. Optional, default TRUE.
thresholdValue	Absolute value to filter out nucleosome calls. It is the minimum number of reads (coverage) in a nucleosome call expressed as reads per million of mapped reads. Optional, default 10.
thresholdPercentage:	Percentile of coverage in the experiment used as threshold to filter out nucleosome calls (i.e., '25%' would mean that only peaks with coverage in the 1st quantile would be considered). Optional, default 35(%)
chr	Chromosome to consider for the analysis in the given input file. By default, all the genomic range is considered. Optional, default NULL.
start	Start genomic position to consider for the analysis in the given input file. By default, all the genomic range is considered. Optional, default NULL.
end	End genomic position to consider for the analysis in the given input file. By default, all the genomic range is considered. Optional, default NULL.
NucDyn	
Parameter	Description
input1, input2	Input BAM from MNase-seq in RData format (from readBAM)
calls1, calls2	Nucleosome calls in GFF format as obtained from nucleR
outputGff	Output of NucDyn in GFF format: position of evictions, inclusions, shifts.
outputBigWig {bw}	Output of NucDyn in BigWig format: -log10 of the p-value of the significance of the differences found.
genome {chrom.sizes}	Chromosome sizes from reference genome

range (All chr chr:start-end)	Genomic range to be analyzed. All: all genome; chr: a whole chromosome; chr:start-end: a specific region given the coordinates. Optional, default All.
plotRData {RData}	Save all the detected changes at the fragment level in RData format for posterior plotting. Optional, default NULL (no RData is saved).
maxDiff	Maximum distance between the centers of two fragments for them to be paired as shifts. (bp) Optional, default 70.
maxLen	This value is used in a preliminar filtering. Fragments longer than this will be filtered out, since they are likely the result of MNase under-digestion and represent two or more nucleosomes (bp). Optional, default 140.
shift_min_nreads	Minimum number of shifted reads for a shift hotspot to be reported {int}. Optional, default 3.
shift_threshold	Threshold applied to the shift hotspots. Only hotspots with a score better than the value will be reported. Notice the score has to be lower than the threshold, since these numbers represent p-values {float}. Optional, default 0.1.
indel_min_nreads	Minimum number of removed/included reads for an insertion or eviction hotspot to be reported {int}. Optional, default 3.
indel_threshold	Threshold applied to the inclusion and eviction hotspots. Only hotspots with a score better than the value will be reported. Notice the score has to be lower than the threshold, since these numbers represent p-values {float}. Optional, default 0.05.
cores	Number of computer threads. Optional, default 1.
equal_size	Trim all fragments to the same size. Optional, default FALSE.
readSize	Length to which all reads will be set in case 'equalSize' is 'TRUE'. It is ignored when 'equalSize' is set to 'FALSE'. Optional, default 140.
NFR	
Parameter	Description
input	Nucleosome calls in GFF format as obtained from nucleR.
output {gff}	Nucleosome Free Regions in GFF format.
minwidth	Minimum length (bp). Optional, default 110bp.
threshold	Maximum length (bp). Optional, default 400bp.
TSS classificaiton	
Parameter	Description
calls	Nucleosome calls as obtained from nucleR. GFF format
genome	Gene positions in the reference genome. GFF Format
output	Classification of TSS according to nucleosome -1 and +1. GFF format.
window	Number of nucleotides on each side of the TSS where -1 and +1 nucleosomes are searched for. Optional, default 300.
open_thresh	Distance between nucleosomes -1 and +1 to discriminate between 'open' and 'close' classes. Optional, default 215.
cores	Number of computer threads. Optional, default 1.
p1.max.downstream	Maximum distance upstream from the TSS to look for +1 nucleosome. Optional, default 20.
Periodicity	
Parameter	Description
calls	Nucleosome calls in GFF format as obtained from nucleR.
reads	Sequence Reads in RData format as obtained from readBAM.
type	Type of reads (single paired)
gffOutput	Periodicity output in GFF format.
bwOutput	Periodicity output in BigWig format.
genes	Position of genes in reference genome. GFF format
chrom_sizes	Chromosome sizes file in the reference genome.
periodicity	Average distance between two consecutive nucleosomes. It is used as the period of the nucleosome coverage signal. It should be defined according to the nucleosome repeat length in the corresponding cell type. Optional, default 165.
cores	Number of computer threads. Optional, default 1.
Stiffness	
Parameter	Description
calls	Nucleosome calls in GFF format as obtained from nucleR.
reads	Sequence data in RData format as obtained from readBAM.
output	Output stiffness for each nucleosome call in GFF format.
range	Genomic range to consider. Format: [str, All chr chr:start-end] where 'All' is all genome, 'chr' is a single chromosome, and 'chr:start-end' is the range indicated by the coordinates. Optional, default 'All'.
t	Temperature (K). Optional, default 310.15.

Table S2. Suggested nucleosome repeat length in different species and cell types.

<i>Cell type/species</i>	<i>Nucleosome Length</i>	<i>Repeat</i>	<i>Source</i>
<i>Drosophila melanogaster</i>	175 bp		(2)
Human CD4+ T cells	180 bp		(3)
<i>Caenorhabditis elegans</i>	175 bp		(4)
<i>Schizosaccharomyces pombe</i>	154 bp		(5)
Mouse ESCs	186 bp		(6)
Mouse NPCs	193 bp		(6)
Mouse MEFs	191 bp		(6)

1. Chereji,R.V., Ramachandran,S., Bryson,T.D. and Henikoff,S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biology*, **19**.
2. Mavrich,T.N., Jiang,C., Ioshikhes,I.P., Li,X., Venters,B.J., Zanton,S.J., Tomsho,L.P., Qi,J., Glaser,R.L., Schuster,S.C., *et al.* (2008) Nucleosome organization in the *Drosophila* genome. *Nature*, **453**, 358–362.
3. Schones,D.E., Cui,K., Cuddapah,S., Roh,T.-Y., Barski,A., Wang,Z., Wei,G. and Zhao,K. (2008) Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, **132**, 887–898.
4. Valouev,A., Ichikawa,J., Tonthat,T., Stuart,J., Ranade,S., Peckham,H., Zeng,K., Malek,J.A., Costa,G., McKernan,K., *et al.* (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, **18**, 1051–1063.
5. Lantermann,A.B., Straub,T., Strålfors,A., Yuan,G.-C., Ekwall,K. and Korber,P. (2010) *Schizosaccharomyces pombe* genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae*. *Nature Structural & Molecular Biology*, **17**, 251–257.
6. Teif,V.B., Vainshtein,Y., Caudron-Herger,M., Mallm,J.-P., Marth,C., Höfer,T. and Rippe,K. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nature Structural & Molecular Biology*, **19**, 1185–1192.

Table S3. Public data available at MuGVRE.

Data	Source
<i>Saccharomyces cerevisiae</i>	
Gene and Gene predictions	Saccharomyces Genome Database ¹
Gene structure / UTRs / transcribed regions	Yassour et al, 2009 ²
Gene Models / introns / 5' 3' UTR's / unannotated transcripts	Nagalakshimi et al. 2008 ³
Transcription Start sites	Zhang, Z and Dietrich FS. 2005 ⁴
Chromatin modifications	Kirmizis A. et al. 2007 ⁵
Nucleosome positions	Mavrich et al. 2008 ⁶
Digital genomic footprinting	Hesselberth et al. 2009 ⁷
H2A.Z nucleosome positions	Albert et al. 2007 ⁸
H2A/H2B, H2A.Z/H2A.Z, H2A.Z/H2B log2 ChIP chip ratio	Guillemette et al. 2005 ⁹
H3K4ac_set1D_on_WT, set1D_H3K4ac_on_H3, WT_H3K4ac_on_H3, WT_H3K4me3_on_H3	Guillemette et al. 2011 ¹⁰
anti-Ac, H2AK7ac, H2BK16ac, H3K14ac, H3K18ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H4K12ac, H4K16ac, H4K5ac, H4K8ac, mock, RNA PolII ChIP_chip	Liu et al. 2005 ¹¹
predicted average nucleosome occupancy, predicted nucleosome potential score, nucleosome sequence read count	Field et al. 2008 ¹²
nucleosome positions, nucleosome signal, nucleosome calling occurrences	Schep et al. 2015 ¹³
ORC, Mcm2p binding, ARS sequences	Xu et al. 2006 ¹⁴
ORC, ARS, Nucleosome positioning	Eaton et al. 2010 ¹⁵
TATA_elements	Rhee and Pugh 2012 ¹⁶
Bur1, Cet1 (Capping enzyme), Ctk1, Elf1, Kin28 (TFIIH), Paf1, Pcf11, Ser2P (RNA Pol II), Ser5P (RNA Pol II), Ser7P (RNA Pol II), Rpb3 (RNA Pol II), Spn1 (lws1), Spt16, Spt4, Spt5, Spt6, Spt6deltaC, Tfg1 (TFIIF), TFIIB	Mayer et al. 2010 ¹⁷
Gal4, Phd1, Rap1, Reb1	Rhee and Pugh 2011 ¹⁸
Nucleosome architecture through cell cycle	Deniz et al. 2016 ¹⁹
Dyad distribution and Normalized nucleosome occupancy profiles from chemical cleavage nucleosome mapping	Chereji et al. 2018 ²⁰
<i>Drosophila melanogaster</i>	
Genes, Transcripts	
Chromatin types through protein binding sites	Filion et al. 2010 ²¹
Nucleosome organization	Mavrich et al 2008 ²²
<i>Homo sapiens</i>	
Refseq Genes	
Ensembl Genes	

1. Saccharomyces Genome Database. Available: <http://www.yeastgenome.org>
2. Yassour M1, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtkoval, Gnirke A, Nusbaum C, Thompson DA, Friedman N, Regev A. Ab initio construction of eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A. 2009. 106(9):3264-9.3.

3. Nagalakshmi U1, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. Thetranscriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008;320(5881):1344-1349.4.
4. Zhang Z, Dietrich FS. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5'SAGE. *Nucleic Acids Res*. 2005. 33(9):2838-515.
5. Kirmizis A, Santos-Rosa H, Penkett CJ, Singer MA, Vermeulen M, Mann M, Bähler J, Green RD,Kouzarides T. Arginine methylation at histone H3R2 controls deposition of H3K4trimethylation. *Nature*. 2007; 449(7164):928-932.6.
6. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF Abarrier nucleosome model for statistical positioning of nucleosomes throughout the yeastgenome. *Genome Res*. 2008 18(7):1073-1083.7.
7. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S,Kuehn MS, Noble WS, Fields S, Stamatoiyannopoulos JA. Global mapping of protein-DNAinteractions in vivo by digital genomic footprinting. *Nat Methods*. 2009. 6(4):283-289.8.
8. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. Translational androtational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome.*Nature*. 2007;446(7135):572-576.9.
9. Guillemette B, Bataille AR, Gévry N, Adam M, Blanchette M, Robert F, Gaudreau L. Varianthistone H2A.Z is globally localized to the promoters of inactive yeast genes and regulatesnucleosome positioning. *PLoS Biol*. 2005; 3(12):e384.10.
10. Guillemette B, Drogaris P, Lin HH, Armstrong H, Hiragami-Hamada K, Imhof A, Bonneil E,Thibault P, Verreault A, Festenstein RJ. H3 lysine 4 is acetylated at active gene promoters andis regulated by H3 lysine 4 methylation. *PLoS Genet*. 2011;7(3):e1001354.11.
11. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol*. 2005; 3(10):e328.12.
12. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E.Distinct modes of regulation by chromatin encoded through nucleosome positioning signals.*PLoS Comput Biol*. 2008; 4(11):e1000216.13.
13. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. Structurednucleosome fingerprints enable high-resolution mapping of chromatin architecture withinregulatory regions. *Genome Res*. 2015; 25(11):1757-1770.14.
14. Xu W, Aparicio JG, Aparicio OM, Tavaré S. Genome-wide mapping of ORC and Mcm2pbinding sites on tiling arrays and identification of essential ARS consensus sequences in *S.cerevisiae*. *BMC Genomics*. 2006 26;7:276.15.
15. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM. Conserved nucleosome positioningdefines replication origins. *Genes Dev*. 2010;24(8):748-75316.
16. Rhee HS, Pugh BF. Genome-wide structure and organization of eukaryotic pre-initiationcomplexes. *Nature*. 2012 18;483(7389):295-301.17.
17. Mayer A, Lidschreiber M, Siebert M, Leike K, Söding J, Cramer P. Uniform transitions of thegeneral RNA polymerase II transcription complex. *Nat Struct Mol Biol*. 2010; 17(10):1272-1278.18.
18. Rhee HS, Pugh BF Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*. 2011 9;147(6):1408-1419.19.
19. Deniz Ö, Flores O, Aldea M, Soler-López M, Orozco M. Nucleosome architecture throughoutthe cell cycle. *Sci Rep*. 2016 28;6:19729.20.
20. Chereji,R.V., Ramachandran,S., Bryson,T.D. and Henikoff,S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biology*, **19**.
21. Filion GJ, van Bommel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B. Systematic protein locationmapping reveals five principal chromatin types in *Drosophila* cells. *Cell*. 2010 15;143(2):212-24.21.
22. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL,Schuster SC, Gilmour DS, Albert I, Pugh BF. Nucleosome organization in the *Drosophila*genome *Nature*. 2008 15;453(7193):358-62

Table S4. Sample statistics per gene computed for nucleR and NucDyn on *S. cerevisiae* cell cycle data from Deniz et al. 2016.

Gene	G1					ND, G1 S			
	Total Nucleosomes	Total Well-Positioned	Total Fuzzy	TSS class	Distance from -1 to +1	Inclusions	Evictions	Shifts+	Shifts-
YKL173W (SNU114 GIN10)	22	10	12	W-open-W	268	0	0	1	0
YKL174C (TPO5)	13	8	5	W-open-W	268	0	0	0	1
YKL175W (ZRT3)	12	9	3	W-open-W	249	0	1	1	1
YKL178C (STE3 DAF2)	11	8	3	W-close-W	168	0	1	0	1
YKL182W (FAS1)	52	11	39	W-open-F	248	0	1	0	1
YKL185W (ASH1)	13	10	3	F-open-W	229	0	2	0	3
YKL187C	16	13	3	W-close-W	161	0	0	1	0
YKL190W (CNB1 CRV1 YCN2)	6	6	0	W-open-W	326	0	0	0	1
YKL192C (ACP1)	4	3	1	W-open-W	265	0	0	1	0
YKL193C (SDS22 EGP1)	7	7	0	F-close-W	204	0	1	1	0
YKL196C (YKT6)	5	5	0	W-open-W	288	0	1	1	0
YKL197C (PEX1 PAS1)	22	18	4	W-close-W	201	0	0	2	0

Table S5. Number of nucleosomes detected by nucleR under different MNase levels.

H4-input	Nucleosomes	H2B-input	Nucleosomes
25U	80160	50U	82543
50U	80666	100U	82102
100U	79366	200U	80188
200U	73689	300U	78614
300U	72775	400U	74275

Table S6. Number of hotspots detected by NucDyn.

	Region	Evictions	Inclusions	Shifts
H2B-input	Genome-wide	3559	278	110
	Promoters	2604	46	30
H4-input	Genome-wide	4072	334	55
	Promoters	3041	39	17

V. Impact of DNA methylation on 3D genome structure

Supplementary Figures

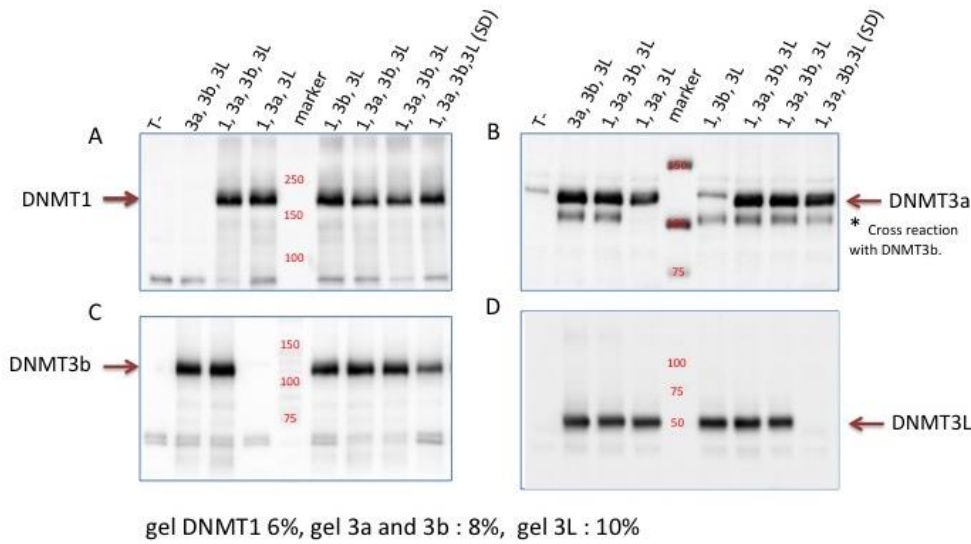


Figure S1. Expression and stability of the 4 DNMTs. (A-D) The lysate from 7 transformants expressing different combinations of the 4 DNMTs, and one control (T-) were loaded in (A) 6% (B,C) 8% and (D) 10% acrylamide gel, transferred onto PVDF membrane and revealed with (A) anti DNMT1 antibody, (B) anti-DNMT3a (C) anti DNMT3b and (D) anti-Flag antibody.

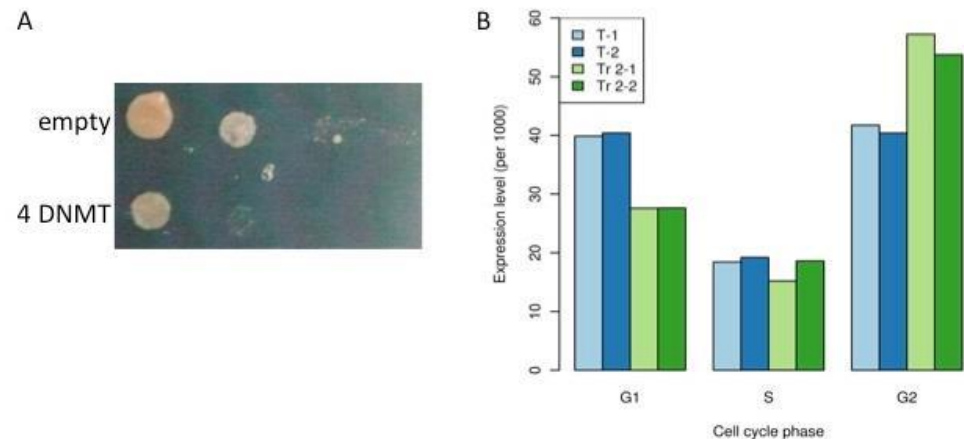


Figure S2. (A) Viability test of one yeast culture transformed with empty vectors (T-1, T-2) and one yeast culture expressing the 4 DNMTs (B) Flow cytometry analysis of two independent yeast cultures transformed with empty vectors (T-1, T-2) and two independent yeast cultures expressing DNMTs (Tr2-1, Tr2-2). Percentage of cells in G2 is larger in methylated samples than in the control samples suggesting a slightly longer G2 phase in the methylated sample.

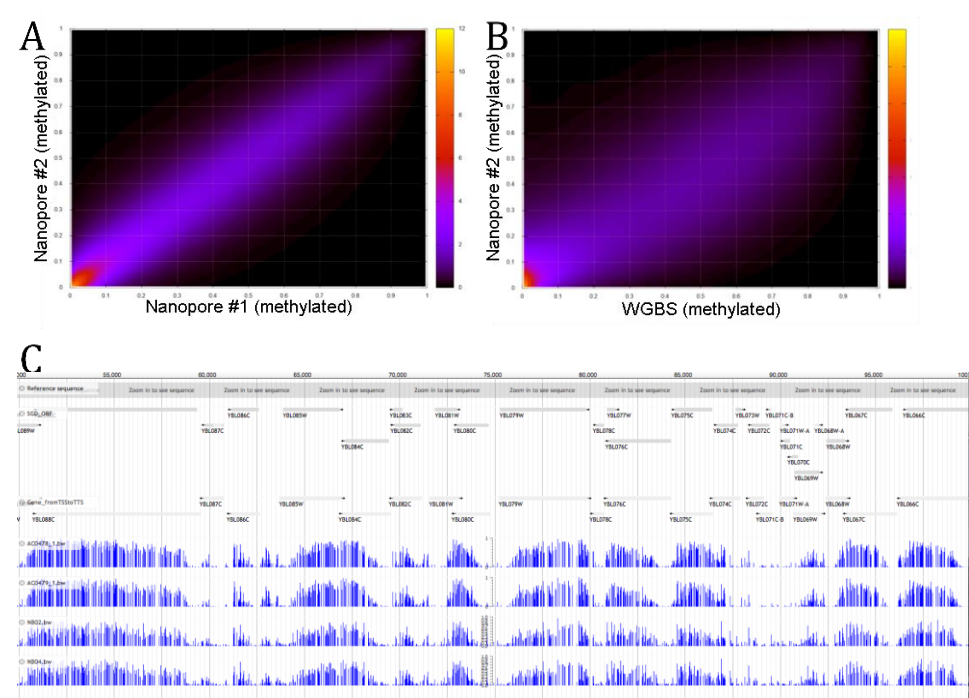


Figure S3. Heatmap showing the pairwise CpG methylation correlation in (A) two nanopore replicates and (B) in one nanopore vs one WGBS replicate. (C) Methylation pattern at the rDNA locus in WGBS (top 2 tracks) and nanopore (bottom tracks) samples for 2 replicas of each condition.

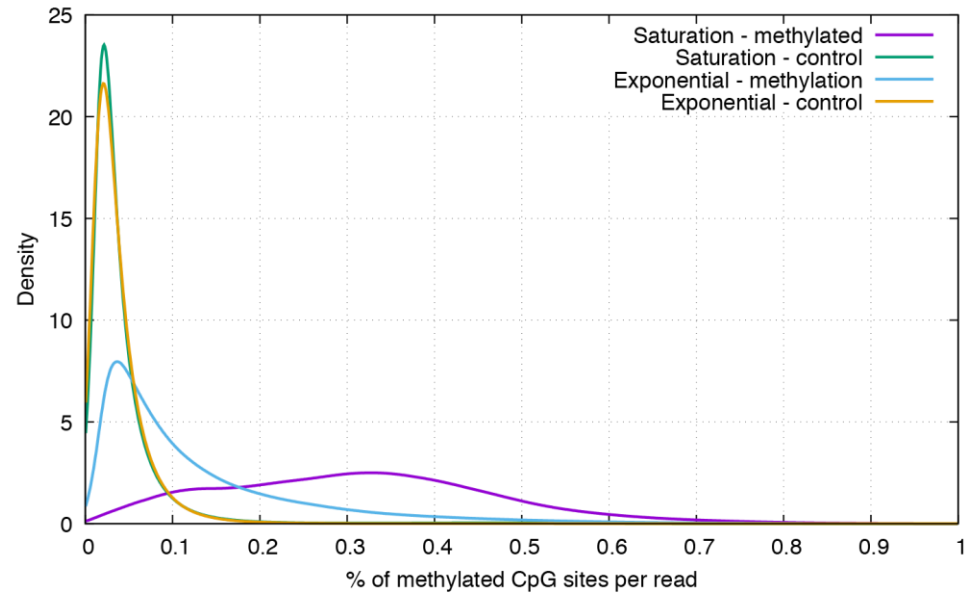
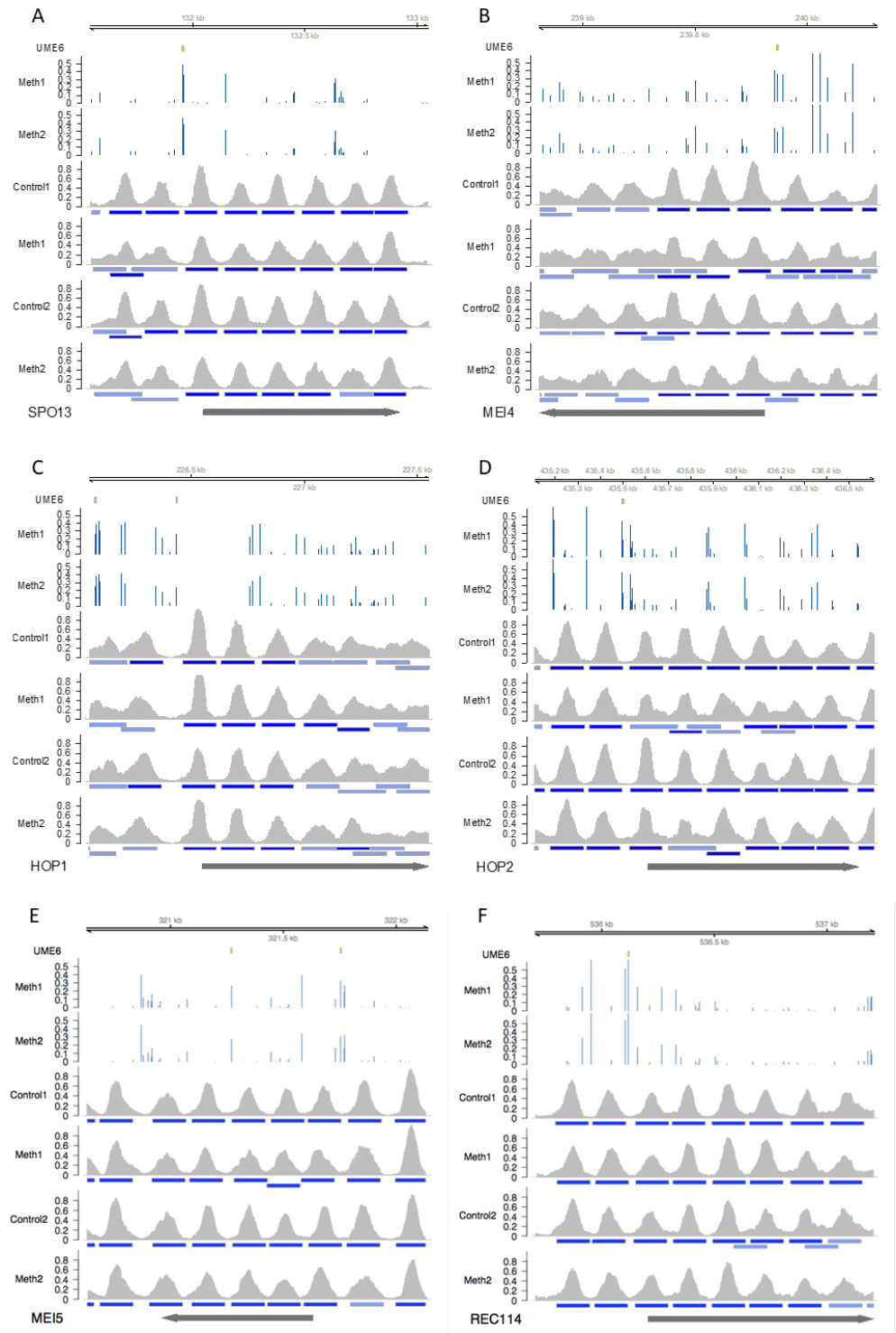


Figure S4. Density estimates of the % of CpG sites methylated per read estimated from Nanopore sequence data for four datasets: methylated cells in stationary phase, control cells in stationary phase, methylated cells in exponential phase and control cells in exponential phase.



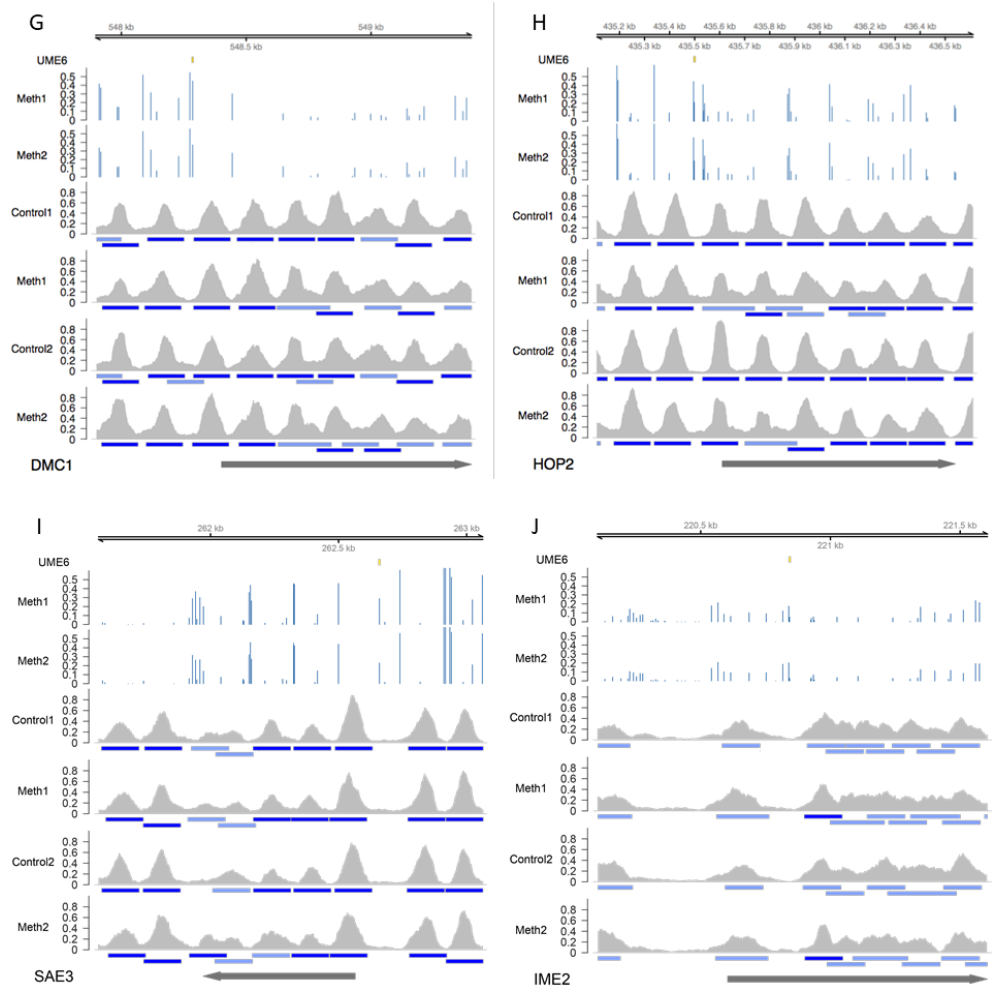


Figure S5. Nucleosome position changes upon DNA methylation for 10 upregulated genes. The position of UME6 DNA binding sequence is indicated as a yellow box, and the methylation levels at individual CpG as a blue histogram for two replicas (Meth1, Meth2). The blue boxes represent the nucleosomes as called by nucleR for the two control and two methylated samples.

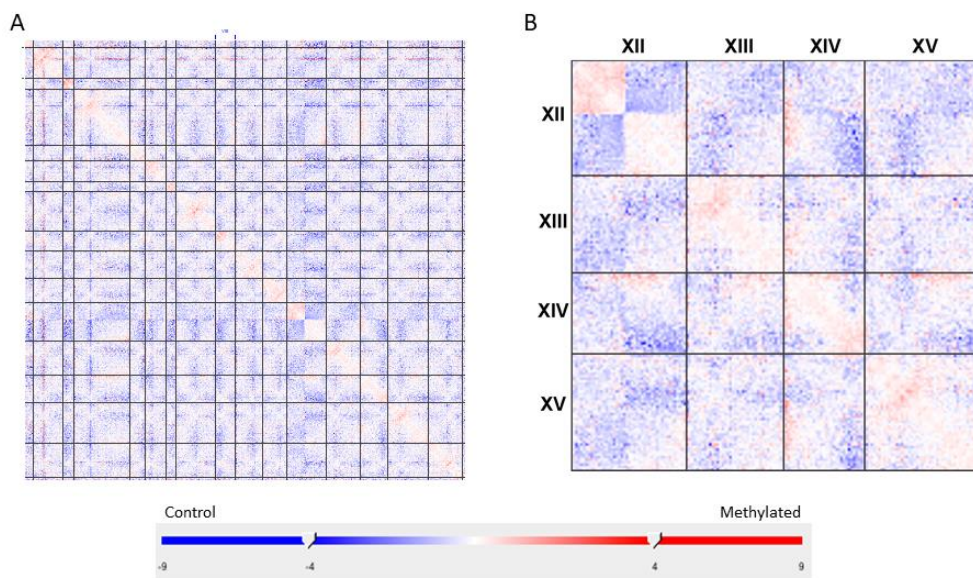
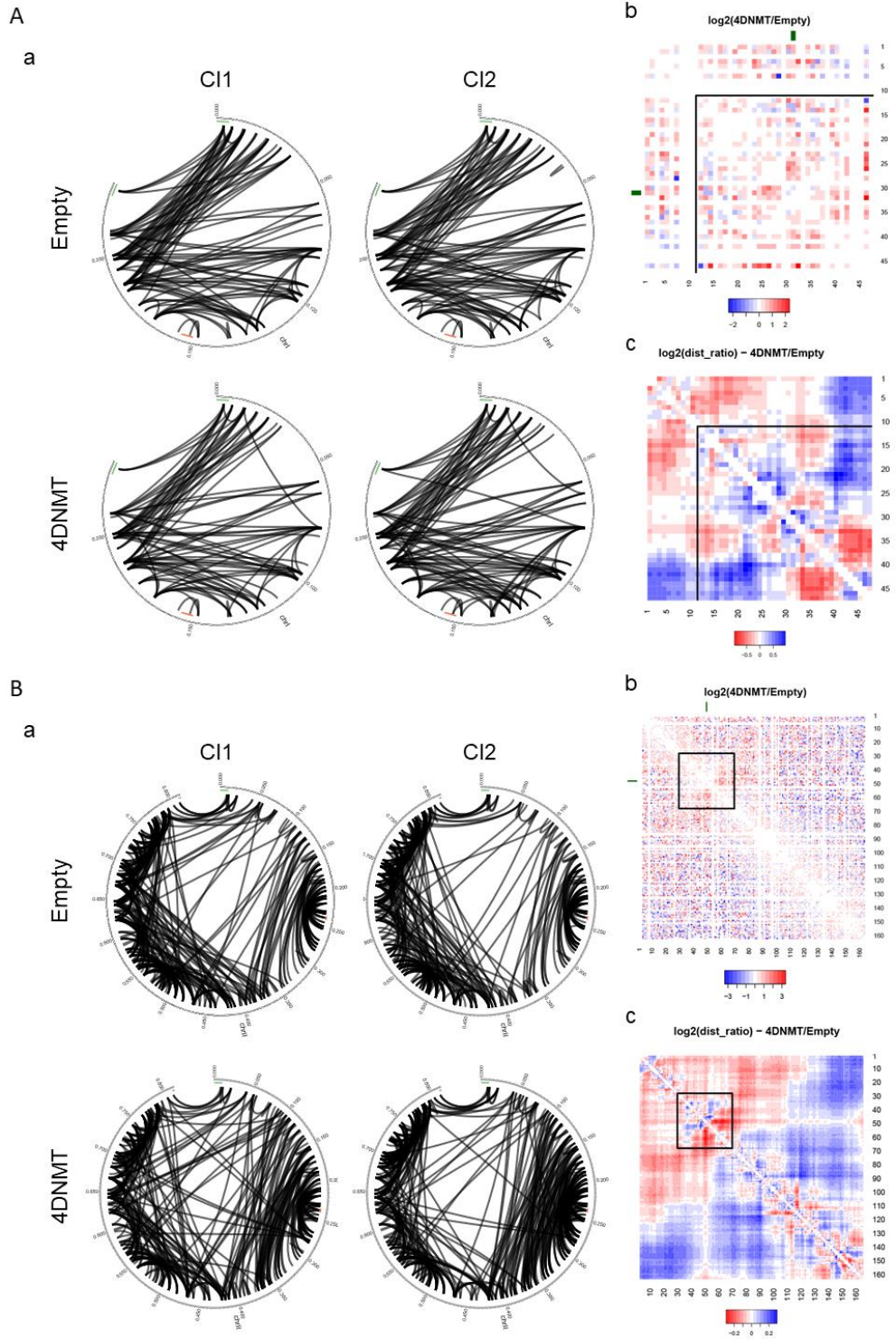
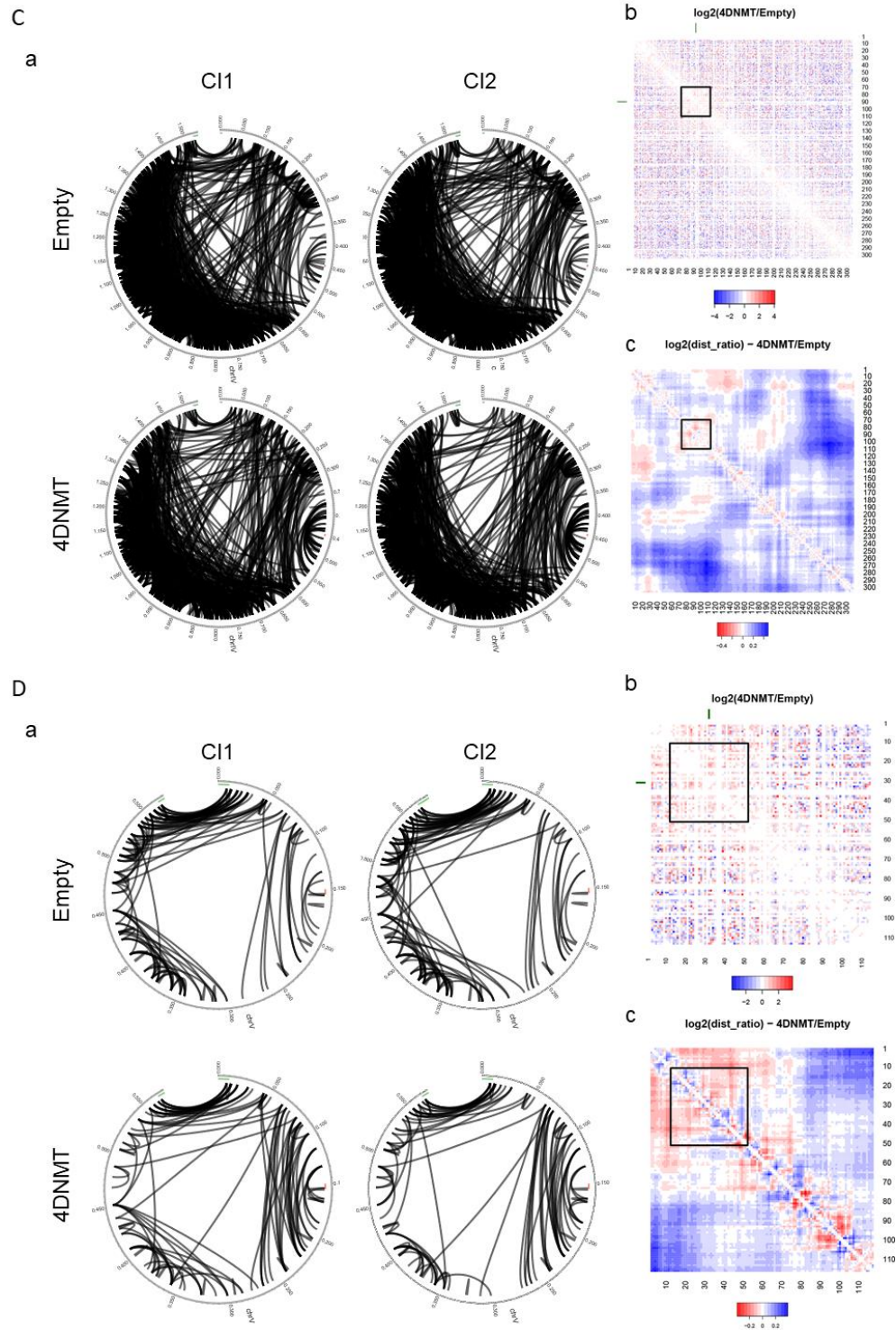
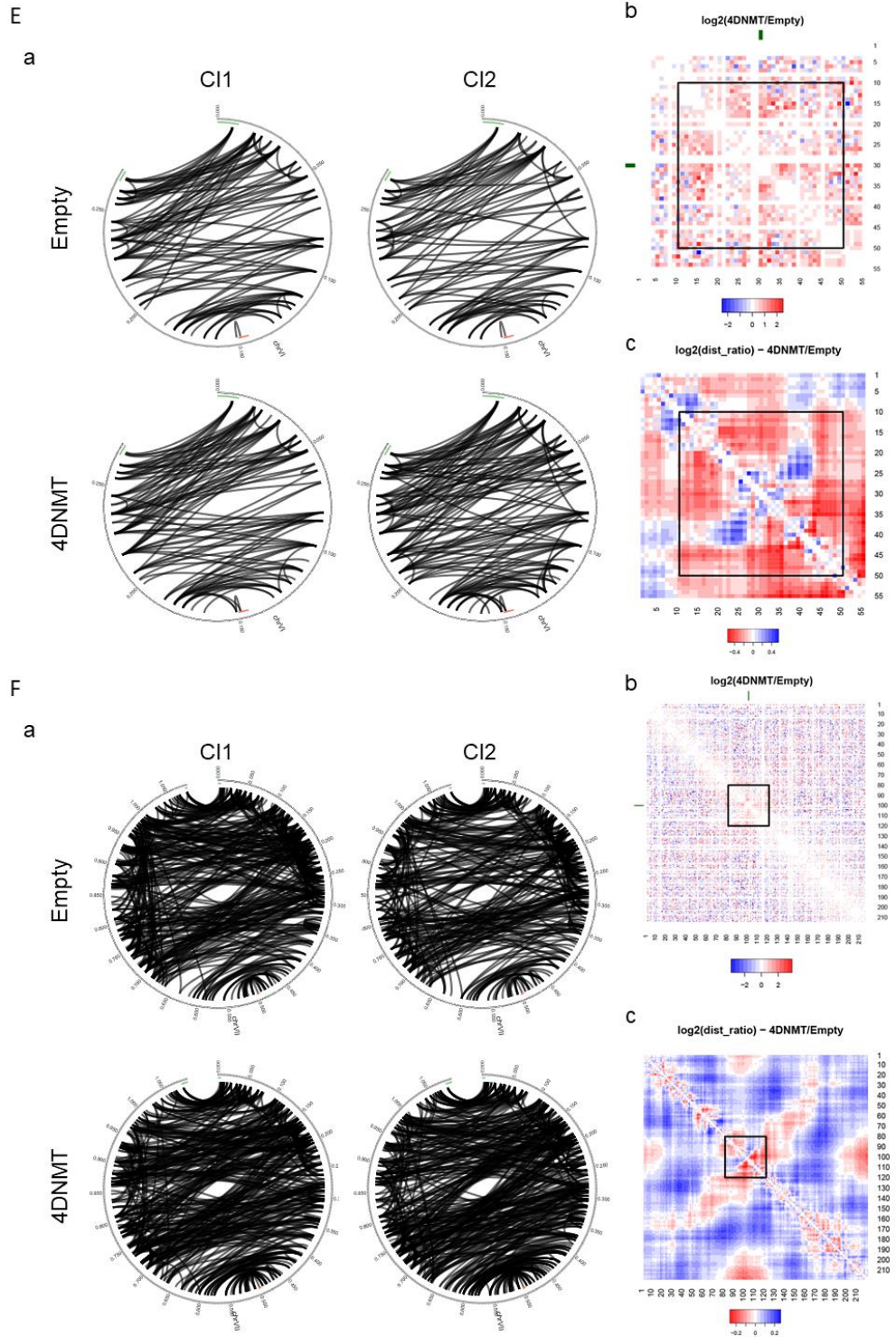
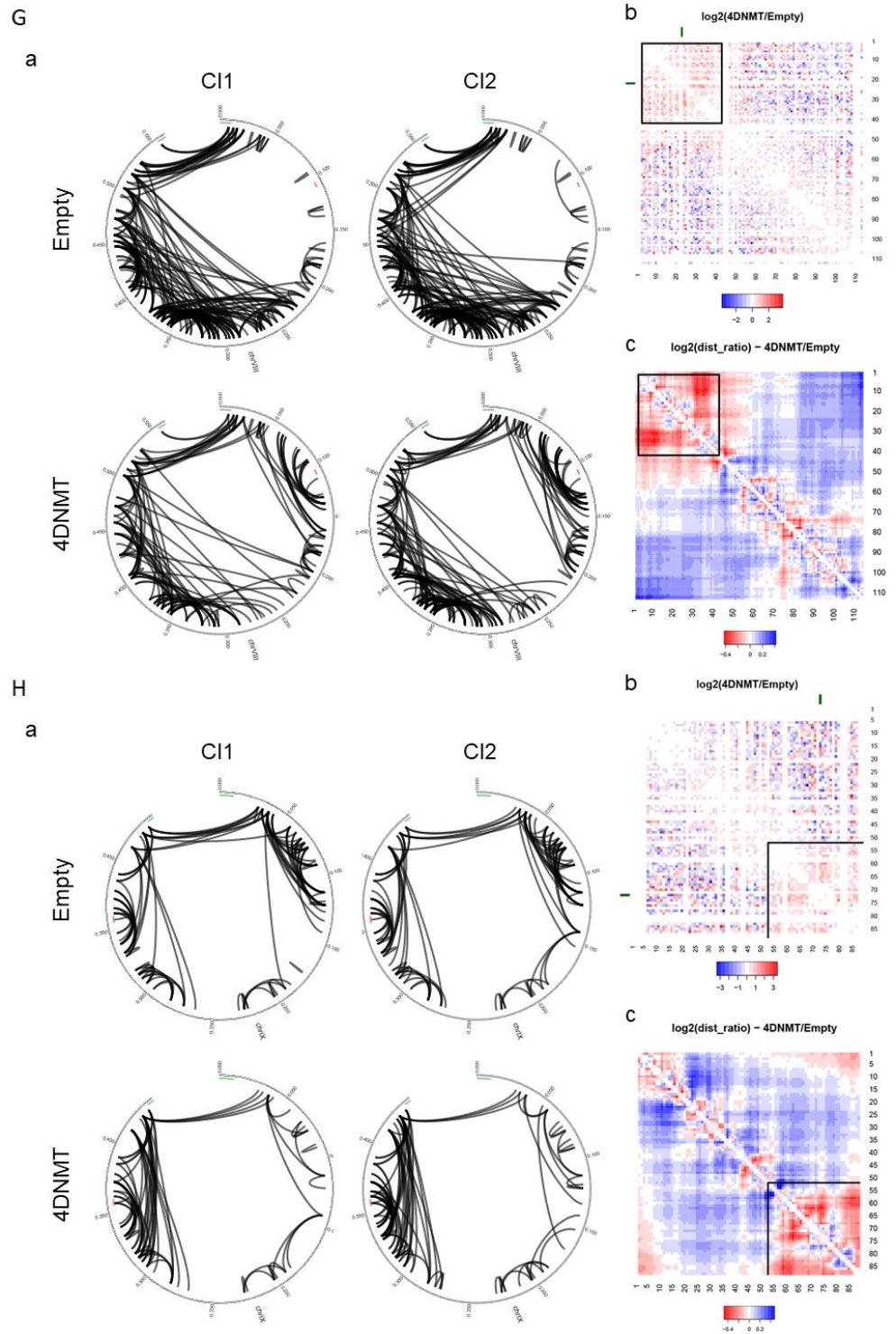


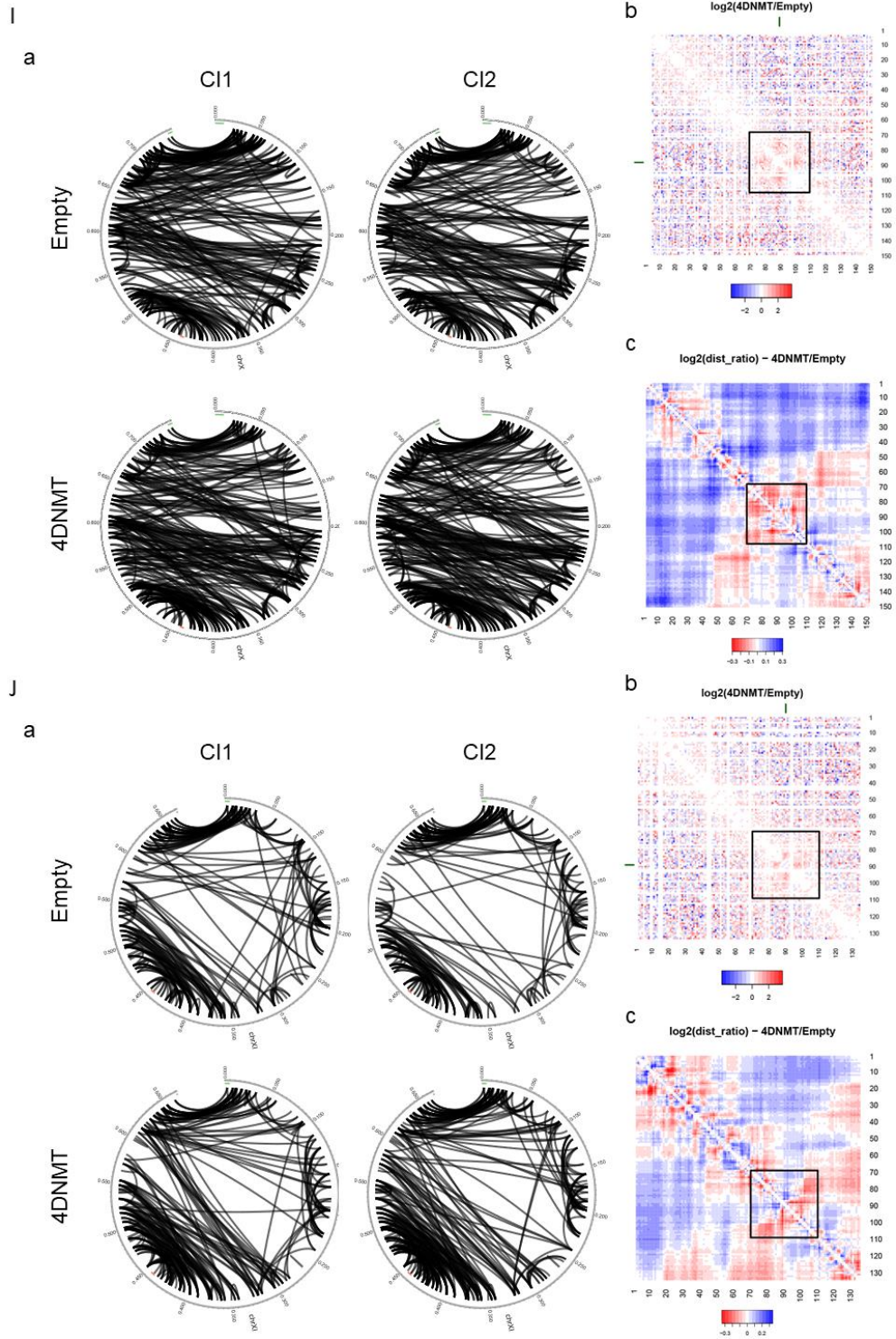
Figure S6. Effect of DNA methylation on 3D genome structure in replica 2. Differential contact frequencies in control and methylation induced samples for (A) whole genome and (B) focus on four chromosomes. Blue indicates interaction with a higher frequency in the non-methylated control sample and red indicates interactions with a higher frequency in the methylated samples.

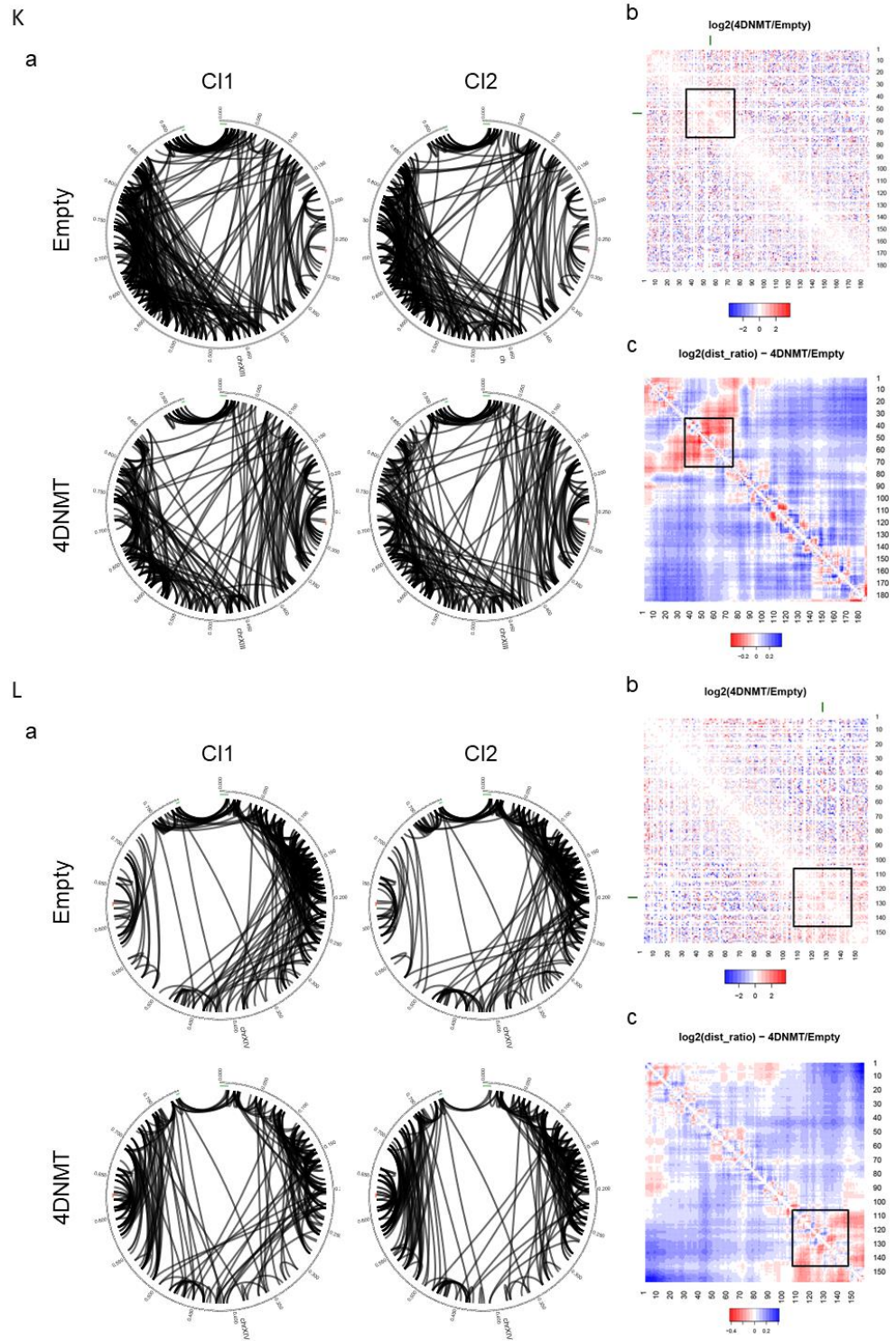












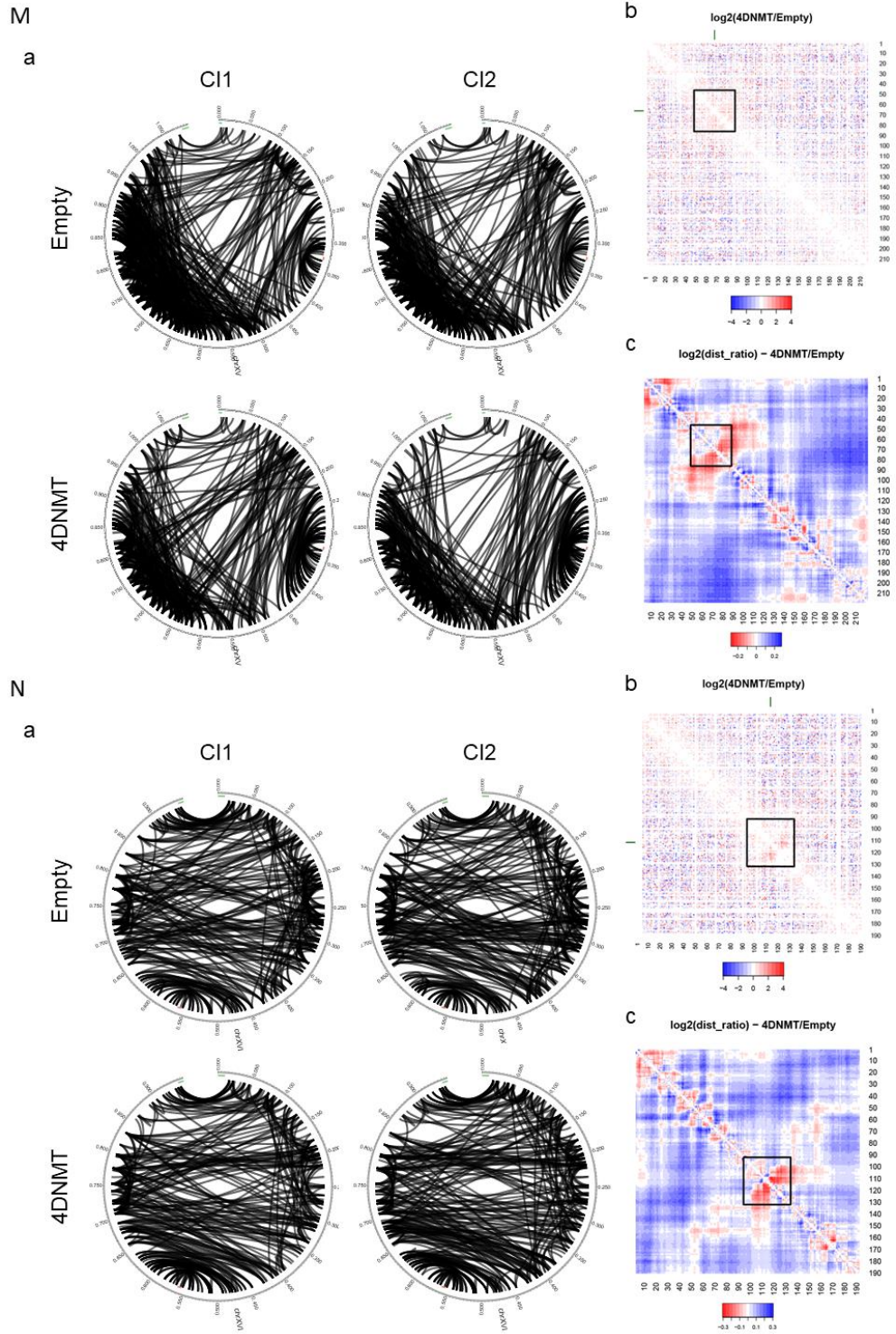


Figure S7. Comparison of interactions in the control and the methylated strains. Panels (A-N) represent each chromosome (chrI, II, IV-XI and XIII-XIV, respectively). (a) Circos diagrams depict each chromosome as a circle. Each arc represents a significant interaction in the control (top) and the methylated sample (bottom) for the two replicas (Cl1, Cl2). The chromosomal position of the centromeres is indicated in red and the telomeres in green. (b) Log2 ratio of the interaction frequencies in the control over the methylated for replica 1. Blue indicates interaction with a higher frequency in the control sample and red indicates interactions with a higher frequency in the methylated sample. (c) Log2 ratio of the distance in the 3D model for the control over the methylated for replica 1. Blue indicates shorter distance in the control sample and red indicates closer in the methylated sample.

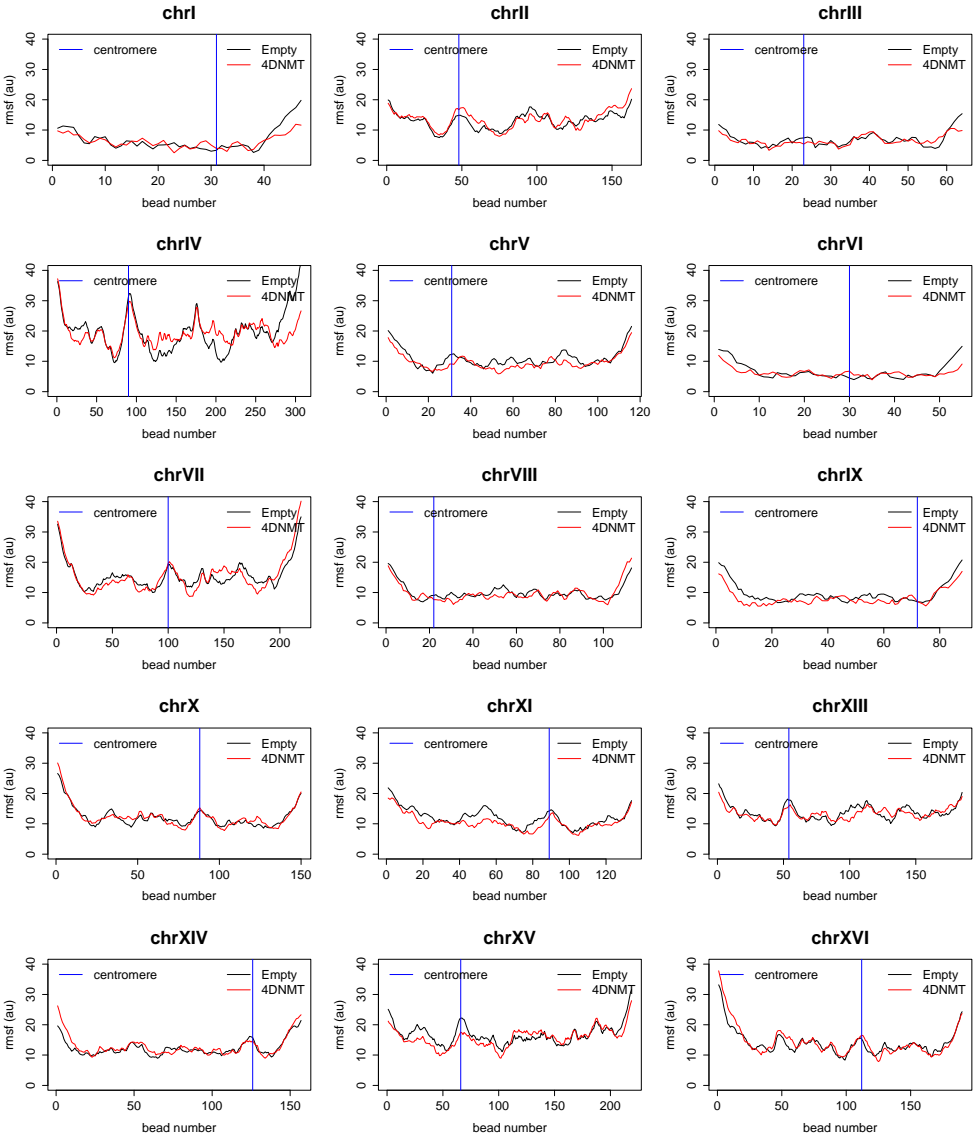


Figure S8. Root mean square fluctuations (RMSF) by bead in each chromosome for the control (black) and methylated (red). The blue line indicates the position of the centromere.

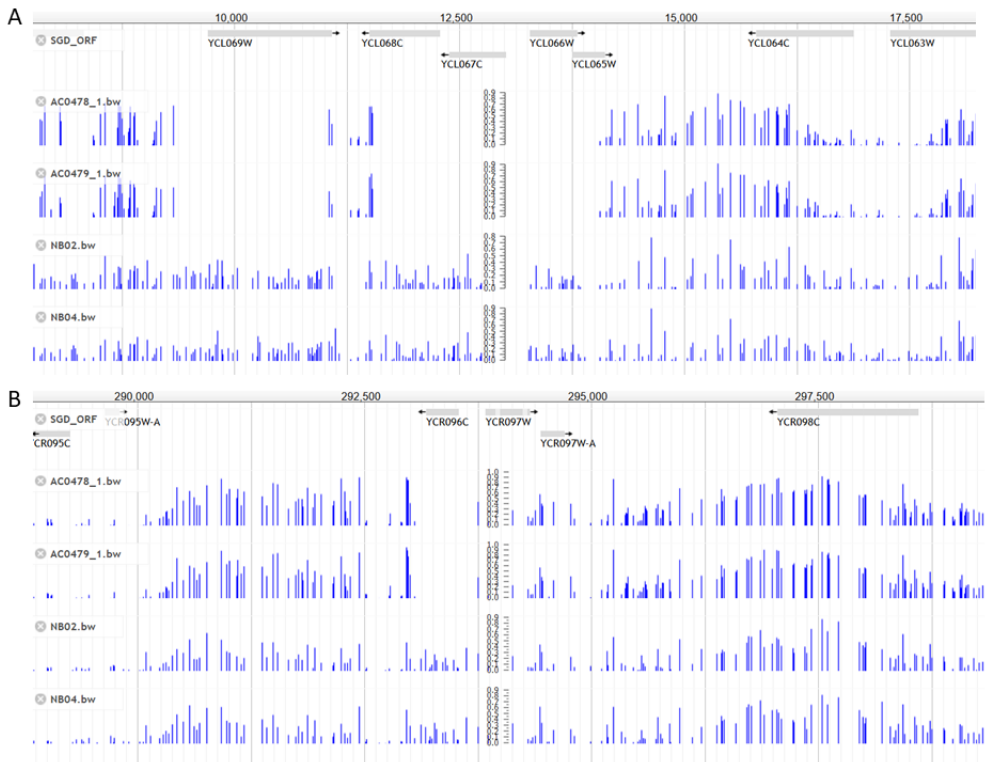


Figure S9. Methylation pattern in WGBS (top 2 tracks) and nanopore (bottom tracks) samples for 2 replicas of each condition at (A) the HML locus, (B) the HMR locus.

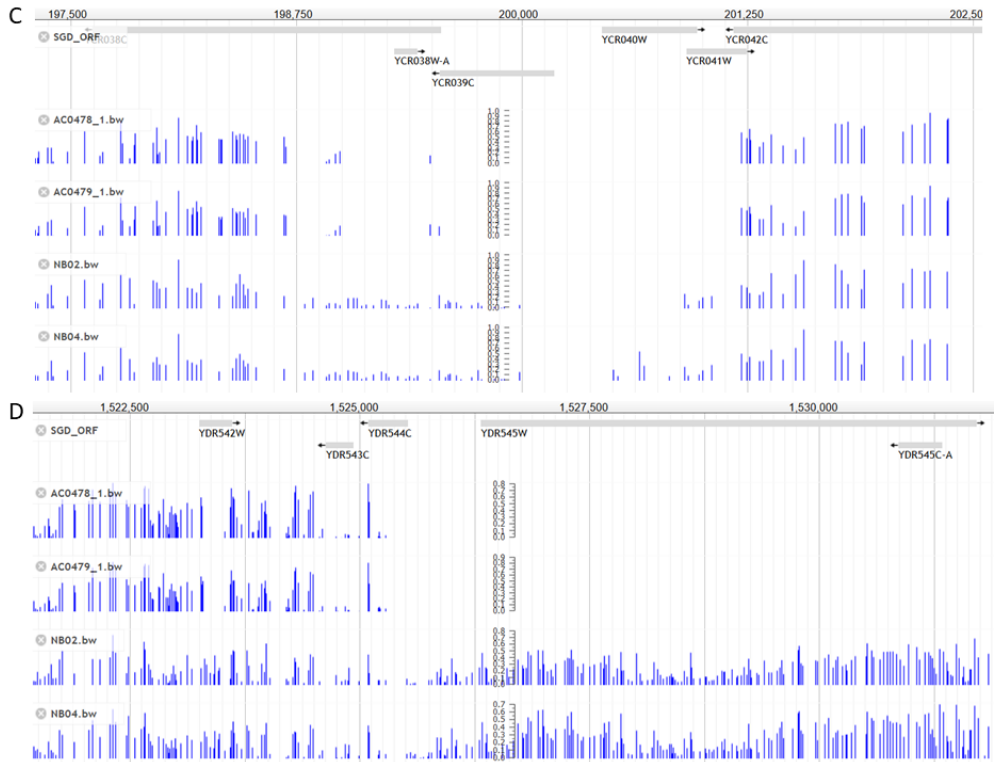


Figure S9 (Cont.). Methylation pattern in WGBS (top 2 tracks) and nanopore (bottom tracks) samples for 2 replicates of each condition at (C) the MAT locus and (D) a telomeric region in chromosome IV.

Suppl. Table S2. Percentage of Fuzzy, Well positioned or not determined nucleosome calls from nucleR on 2 control and 2 methylated replicas in saturation.

Sample	Fuzzy	Well positioned	Not determined	total
Ctrl Rep1	34759 (47.36%)	38320 (52.22%)	307(0.42)	73386
Ctrl Rep2	36322(49.64%)	36544 (49.95%)	298 (0.41%)	73164
Met Rep1	39400 (53.12%)	34362 (46.33%)	403 (0.54%)	74165
Met Rep2	38222 (51.61%)	35425 (47.83%)	416 (0.56%)	74063

Suppl. Table S3. Expression changes and URS1 methylation levels of a subset of early meiotic genes

Name gene	Gene ID	Samples in G1			Samples at saturation		
		Differential expression LOG2FC	p-adj	Methylation level at URS1 site	Differential expression LOG2FC	p-adj	Methylation level at URS1 site
SAE3	YHR079C	6.20	3.38E-21	0.14-0.145	7.43	7.21E-04	0.404-0.37
MEI5	YPL212C	3.84	1.07E-04	0.354-0.318	6.98	1.44E-03	0.509-0.447
GMC2	YLR445W	3.69	5.41E-02	0.176-0.193	6.94	5.98E-03	0.414-0.494
HOP2	YGL033W	2.42	4.26E-03	0.074-0.074	6.83	1.17E-03	0.315-0.392
HED1	YDR014W-A	2.98	3.09E-17	0.203-0.139	4.97	7.21E-04	0.418-0.339
SPO13	YHR014W	4.06	8.93E-07	0.27-0.205	4.95	2.24E-03	0.536-0.554
				0.211-0.181			0.386-0.406
MEK1	YOR351C	3.23	1.42E-02	0.354-0.297	4.89	5.47E-03	0.517-0.476
REC114	YMR133W	3.05	2.32E-2	0.372-0.35	4.84	4.00E-03	0.798-0.811
MER1	YNL210W	2.65	3.06E-01	0.368-0.401	4.75	5.65E-02	0.716-0.749
SPO11	YHL022C	3.12	1.44E-11	0.196-0.222	3.76	3.19E-03	0.727-0.69
DMC1	YER179W	3.08	9.06E-13	0.312-0.212	2.97	1.84E-03	0.469-0.443
	YER044C-A						
MEI4	A	2.19	4.61E-03	0.165-0.13	2.74	1.52E-02	0.414-0.465
HOP1	YIL072W	1.70	1.34E-03	0.088-0.034	2.55	3.19E-03	0.301-0.262
ZIP1	YDR285W	1.28	1.01E-01	0.125-0.104	2.46	5.21E-03	0.348-0.321
REC102	YLR329W	1.69	2.44E-02	0.311-0.197	2.29	3.66E-08	0.533-0.419
IME2	YJL106W	-0.06	2.60E-01	0.011-0.011	2.27	6.04E-03	0.07-0.08
SPO16	YHR153C	0.65	6.16E-01	0.02-0.043	2.04	1.36E-02	0.224-0.155
REC104	YHR157W	-0.54	4.09E-01	0.039-0.033	1.15	2.50E-02	0.114-0.05
RED1	YLR263W	-0.79	1.87E-02	ND	0.18	5.88E-01	ND
RIM4	YHL024W	2.87	4.84E-17	0.012-0	-0.12	8.22E-01	0.082-0.048

Suppl. Figure S4. Number of reads filtered out in the Hi-C processing with TADbit.

	Empty-CI1		Empty-CI2		4DNTM-CI1		4DNMT-CI2	
	Number of reads	(%)	Number of reads	(%)	Number of reads	(%)	Number of reads	(%)
Total reads	77 290 108		84 096 809		63 980 708		60 623 554	
Mapped both	67 901 532	100.00%	67 901 533	100.00%	55 086 767	100.00%	53 623 524	100.00%
1- self-circle	1 748 467	2.58%	2190336	2.96%	1 623 463	2.95%	1291620	2.41%
2- dangling-end	9 508 945	14.00%	9601236	12.98%	9 017 932	16.37%	6208357	11.58%
3- error	539 769	0.79%	544375	0.74%	492 032	0.89%	552491	1.03%
4- extra dangling-end	17 466 889	25.72%	19224967	25.99%	14 483 116	26.29%	15732626	29.34%
5- too close from RES	8 959 962	13.20%	9812011	13.26%	7 056 667	12.81%	7737173	14.43%
6- too short	1 021 588	1.50%	1120886	1.52%	833 844	1.51%	918430	1.71%
7- too large	8 551 118	12.59%	9156511	12.38%	7 686 856	13.95%	6982601	13.02%
8- over-represented	5 154 043	7.59%	5511705	7.45%	4 950 744	8.99%	4553967	8.49%
9- duplicated	31 496 318	46.39%	41497816	56.09%	20 691 442	37.56%	20287110	37.83%
10- random breaks	4 162 523	6.13%	4363097	5.90%	4 007 872	7.28%	3326301	6.20%
Filtered reads	28 515 650	42.00%	25 653 808	37.78%	25 871 440	46.96%	27 031 064	50.41%